UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GIOVANNI APARECIDO DA SILVA OLIVEIRA

COMPUTAÇÃO NA NUVEM

Uma Visão Introdutória com Ênfase no Modelo de *Deployment* Público

RIO DE JANEIRO
2021

GIOVANNI APARECIDO DA SILVA OLIVEIRA

COMPUTAÇÃO NA NUVEM

Uma Visão Introdutória com Ênfase no Modelo de *Deployment* Público

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Profª Dra. Silvana Rossetto

Co-orientador:

RIO DE JANEIRO

2021

CIP - Catalogação na Publicação

O48c

Oliveira, Giovanni Aparecido da Silva
COMPUTAÇÃO NA NUVEM: uma visão introdutória com ênfase no Modelo de Deployment Público / Giovanni Aparecido da Silva Oliveira. -- Rio de Janeiro, 2021.
88 f.

Orientadora: Silvana Rosetto.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciência da Computação, 2021.

1. Computação na nuvem. 2. Sistemas distribuídos. 3. Modelos de computação. 4. Provedores públicos. I. Rosetto, Silvana, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

GIOVANNI APARECIDO DA SILVA OLIVEIRA

COMPUTAÇÃO NA NUVEM
Uma Visão Introdutória com Ênfase no Modelo de *Deployment* Público

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciência da Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 09 de julho de 2021

BANCA EXAMINADORA:

Profa. Silvana Rossetto, D.Sc.
(Instituto de Computação - UFRJ)

Prof. Felipe Acker, Ph.D.
(DMA - Instituto de Matemática - UFRJ)

Profa. Maria Luiza Machado Campos, Ph.D.
(Instituto de Computação - UFRJ)

Profa. Valéria Menezes Bastos, D.Sc.
(Instituto de Computação - UFRJ)

Dedico esse trabalho aos meus pais, Carlos Alberto de Oliveira e Jovana Mori da Silva, e avó, Yolanda Mori da Silva, que muito contribuíram para minha formação acadêmica, moral e espiritual.

AGRADECIMENTOS

Ao Pai, ao Filho e ao Espírito Santo.

"Was erschrickst du desshalb?
- Aber es ist mit dem Menschen wie mit dem Baume.
Je mehr er hinauf in die Hoehe und Helle will,
um so staerker streben seine Wurzeln erdwaerts,
abwaerts, in's Dunkle, Tiefe,
- in's Boese.
... Also sprach Zarathustra"

Friedrich Wilhelm Nietzsche

RESUMO

Desde o surgimento do primeiro provedor público em 2006, o modelo de computação na nuvem tem tornado-se extremamente popular entre os consumidores de serviços de tecnologia da informação. Esse trabalho provê uma visão geral sobre a computação na nuvem, passando antes por uma análise das tecnologias que viabilizaram esse modelo. Posteriormente, o trabalho realiza uma comparação entre os provedores públicos de maior relevância na atualidade (AWS, GCP e Azure), baseando-se em dados obtidos a partir de referências bibliográficas. O resultado da comparação permitiu o estabelecimento de um ranqueamento parcial entre os provedores, de acordo com a característica analisada. A avaliação do desempenho da infraestrutura de rede permitiu identificar problemas isolados e locais na infraestrutura intra e inter-regional dos provedores. Estatísticas de uso e experimentação de diferentes provedores, dentro ambiente empresarial, indicam uma possível intensificação no uso *deployments* multi-nuvem e híbridos em sistemas corporativos.

Palavras-chave: computação na nuvem; sistemas distribuidos.

ABSTRACT

Since the emergence of the first public provider, the cloud computing model is becoming extremely popular between IT services consumers. This work builds an overview about cloud computing, starting by reviewing the technologies that made it feasible. In the second part, this work compares the currently most relevant public cloud providers (AWS, GCP and Azure), based on data presented by referenced works. The results allow us to establish a partial order for ranking the providers, according to each analysis criteria. It was possible to identify isolated and localized issues in providers' intra and inter-region, based benchmarks for network performance. Statistics of provider usage inside the corporate environment show a potential increase of multicloud and hybrid deployments.

Keywords: cloud computing; distributed systems.

SUMÁRIO

1 INTRODUÇÃO

Em 1961, em seu discurso durante o evento de comemoração do centenário do Instituto de Tecnologia de Massachusetts, John McCarthy introduziu a ideia da "computação utilitária", onde serviços de computação seriam oferecidos como utilidade pública (assim como a luz elétrica).

Quase 40 anos depois, a computação utilitária rematerializa-se na forma de computação na nuvem, construída por meio da combinação das tecnologias de:

- **Redes de computadores:** que tornam possível a comunicação entre os diversos elementos computacionais de sistemas distribuídos em geral;
- **Computação em grade:** que torna possível a submissão e execução de cargas de trabalho, partindo de diversos clientes diferentes simultaneamente, sobre uma coleção de recursos computacionais homogenizados, unificados e federados;
- **Virtualização:** que torna possível o compartilhamento da mesma infraestrutura física e homogenizada de hardware, por diversos consumidores, potencialmente distintos, de forma isolada e simultânea.

A popularização da Internet de banda-larga viabilizou a implementação do modelo de *deployment* público, onde os serviços na nuvem são oferecidos de forma comercial e os consumidores advêm de diferentes instituições independentes entre sí. O marco inicial da disponibilização de serviços na nuvem sobre esse modelo de *deployment* se deu com o lançamento do provedor Amazon Web Services (AWS), em 2006, oferecendo Infraestrutura como um Serviço (IaaS) na nuvem.

Com o tempo, mais provedores surgiram oferecendo diferentes serviços de computação na nuvem e esse modelo de computação se tornou extremamente popular. Atualmente, o mercado de computação na nuvem movimenta centenas de bilhões de dólares anualmente e tem crescido monotonamente há, pelo menos, quatro anos.

O objetivo desse trabalho é oferecer ao leitor uma visão geral e introdutória da computação na nuvem, com foco no modelo de *deployment* público, com base em um amplo levantamento bibliográfico e provendo uma conexão com o cenário contemporâneo, por meio da comparação entre os provedores públicos mais populares da atualidade.

A definição do conjunto de protocolos que fundamenta a Internet é originalmente dada por (CERF; KAHN, 1974) e uma síntese histórica rica de seu desenvolvimento, dada por (LEINER et al., 2009). A computação em grade é definida em (FOSTER; KESSELMAN, 2003) e (FOSTER, 2002), uma síntese histórica é dada por (FOSTER; KESSELMAN, 2011). No caso da virtualização, uma definição e síntese histórica são apresentados em (CAMPBELL; JERONIMO, 2006).

Em (MELL; GRANCE et al., 2011) é realizada a definição da computação na nuvem de forma sintética e objetiva. O trabalho (LIU et al., 2011) realiza uma análise mais profunda sobre a arquitetura e outros aspectos conceituais da computação na nuvem. Em (BLOKDIJK; MENKEN, 2009) e (RUPARELIA, 2016), o modelo de computação na nuvem é apresentado de forma didática, expandindo as definições originais e contemplando a história do modelo. Uma visão geral do modelo de computação na nuvem é apresentada em (PRAJAPATI; SHARMA; BADGUJAR, 2018) e (RIMAL; CHOI; LUMB, 2009). Em (PENG et al., 2009) é apresentada uma comparação de diferentes plataformas de computação na nuvem. Em termos gerais, o presente trabalho pode ser visto como uma atualização analítica de (RIMAL; CHOI; LUMB, 2009), utilizando definições mais bem consolidadas, em decorrência do amadurecimento da computação na nuvem, e com caráter mais didático, provendo uma contextualização a partir da apresentação das tecnologias identificadas como fundamentais para o modelo de computação na nuvem.

A escolha das referências bibliográficas se deu de acordo com a completude, clareza e coerência do conteúdo apresentado; de acordo com a fundamentação acadêmica do trabalho, avaliada através da diversidade e qualidade das referências utilizadas pelo trabalho; e de acordo com o impacto acadêmico da publicação, avaliado por meio do número de trabalhos acadêmicos que fazem referência ao trabalho em questão. A respeitabilidade do veículo original de publicação do trabalho também foi considerada. Inevitavelmente, principalmente no caso de atualidades, há um grande número de referências ao conteúdo da Web. Para esses casos, procurou-se utilizar informações publicadas sobre o domínio de autoridades em computação na nuvem como AWS, RedHat e GCP. A análise da computação na nuvem e suas tecnologias fundamentais será feita seguindo uma estrutura comum: uma definição sucinta, a extensão dessa definição, uma revisão histórica e um resumo de atualidades.

A fim de pragmatizar o conhecimento obtido com os capítulos teóricos, apresenta-se uma comparação entre os principais provedores de computação na nuvem no modelo de *deployment* público. Essa comparação considera os aspectos de mercado, popularidade, oferta de serviços, disposição geográfica da infraestrutura de nuvem, disponibilidade (avaliada como produto da contagem de *outages* dos últimos 365 dias e de oferta de Acordos de Níveis de Serviço (SLAs)), *benchmarks* de desempenho da infrestrutura de rede, tarifação (avaliada como resultado da comparação entre o valor de tarifação obtido para um caso de uso considerando a utilização de recursos computacionais, de armazenamento e de rede, com características similares) e as certifações das quais dispõem os provedores.

A análise dos *benchmarks* de desempenho de rede basearam-se em dados obtidos em (THOUSANDEYES, 2019-2020). A complexidade associada à reprodução dos experimentos apresentados por essa referência, inviabiliza a realização dos mesmos em decorrência do distanciamento do escopo principal do presente trabalho. Por esse motivo, utilizou-se o conjunto de dados obtidos empiricamente pela referência.

Sumarizando: No capítulo 2, é realizada uma revisão histórica e conceitual das tecnologias precursoras que tornaram possível a implementação da computação na nuvem. No capítulo 3, é realizada uma conceituação geral das características essenciais, modelos de serviço e modelos de *deployment* da nuvem. No capítulo 4, é apresentada uma síntese histórica de alguns dos marcos mais relevantes para a realização e popularização do modelo de computação na nuvem. No capítulo 5, é realizada uma comparação entre os serviços oferecidos pelos principais provedores públicos da atualidade. O capítulo 6 discute os resultados obtidos neste trabalho.

2 FUNDAMENTAÇÃO

Embora só tenha ganhado notoriedade nas últimas décadas (AWS, 2006a)(GRIFFIN, 2018), a ideia de realizar computação em infraestruturas públicas, com características similares ao que conhecemos hoje por computação na nuvem, já existia, pelo menos, desde os anos 60 (GREENBERGER, 1964)(GARFINKEL, 1999).

Nesse interim de mais de 40 anos entre as primeiras idealizações e a consolidação dos provedores de nuvem pública, desenvolveram-se as tecnologias que tornaram essa empreitada possível. Identifica-se neste trabalho como os três pilares da computação na nuvem: redes de computadores, computação em grade e virtualização (REDHAT, 2021d)(REDHAT, 2021e)(LIU et al., 2011)(GOYAL; DADIZADEH, 2009).

As próximas seções deste capítulo revisam de forma breve a história das tecnologias que tornaram possível a realização do modelo de computação na nuvem. A revisão das redes de computadores, se dá sob o escopo restrito da Internet, em decorrência da extensão do tema. A Internet é especialmente importante para os modelos de *deployment* híbrido e público (ver Seção 3.4), sendo o modelo de deployment público foco deste trabalho.

2.1 INTERNET

O estabelecimento de mecanismos de comunicação em rede é fundamental, não apenas para a computação na nuvem, mas para sistemas distribuídos, em geral (TANENBAUM; STEEN, 2007). Nessa seção é apresentada uma revisão histórica da Internet, que desempenha um papel fundamental principalmente em *deployments* públicos e híbridos.

O conteúdo apresentado nesta seção, baseia-se principalemnte no artigo original de definição dos protocolos da Internet (CERF; KAHN, 1974) e em um artigo contendo o relato histórico da Internet do ponto de vista de nove pesquisadores que estiveram intimamente relacionados com o desenvolvimento da tecnologia (LEINER et al., 2009) .

2.1.1 Conceituação

Definição 1 ***Internet:*** *Sistema global de redes de computadores interconectadas com governança decentralizada. Caracteriza-se pelo uso de infraestruturas públicas de comunicação e pelo uso da pilha de protocolos TCP/IP.*

2.1.2 Pré-história

Em 1962, J.C.R. Licklider publicou pelo MIT uma série de memorandos que discorrem sobre o seu conceito de "Rede Galática". Ele vislumbrou uma rede global de computadores por meio da qual os usuários obteriam acesso a dados e programas, independente da

localização. No mesmo ano, Licklider se tornou o primeiro líder do grupo de pesquisa, o DARPA, que foi um dos grupos mais proeminentes na conceituação inicial de redes de computadores.

Ainda no MIT, Leonard Kleinrock publicou em 1964 o primeiro livro sobre comutação de pacotes (KLEINROCK, 1964), completando o trabalho que iniciara com a publicação de seu artigo, sobre o mesmo tema, em 1961 (KLEINROCK, 1961). A escolha pela comutação de pacotes em detrimento da comutação de circuitos foi uma decisão fundamental que modelou a Internet e permanece até os dias de hoje.

As primeiras experimentações documentadas da teoria de redes recém surgida se deram no ano seguinte, 1965, quando Thomas Merrill e Lawrence Roberts estabeleceram conexão entre um computador TX-2 em Mass (MI) e um Q-32 na Califórnia (MARILL; ROBERTS, 1966). O meio utilizado foi uma linha telefônica dial-up de baixa velocidade. Com isso, estabeleceu-se a primeira Rede de Longa-Distância (WAN) conhecida. O experimento permitiu constatar a inadequação do sistema telefônico de comutação de circuitos para realizar comunicações desse tipo.

Em 1966, Roberts foi para a DARPA onde elaborou conceitos essenciais de redes de computadores e onde deu início à implementação da ARPA Net.

Em 1967, Roberts pubica um artigo (ROBERTS, 1967), onde define a ARPA Net como uma rede experimental de computadores, interligando computadores de diversos grupos de pesquisa dos Estados Unidos. Um dos objetivos da ARPA Net era promover infraestrutura e protocolo unificados, para viabilizar a comunicação entre os computadores presentes na rede.

Nessa época, começavam a se proliferar os protocolos de rede. Para facilitar a colaboração dos pesquisadores que estavam envolvidos com o desenvolvimento desses protocolos abertos, criou-se o *Request for Comments* (RFC). Nessas publicações, os autores submetem propostas de protocolos relevantes e a comunidade colabora com comentários, adições e correções. Embora tenha surgido com um propósito informal, as RFCs desempenharam um papel fundamental na colaboração e difusão de padrões e algoritmos não proprietários. Atualmente, existem mais de 9.000 RFCs com diversos status para indicar o grau de maturidade de cada uma das propostas.

Em dezembro de 1970, Steve Crocker publicou um protocolo de comunicação ponto-a-ponto entre computadores, chamado *Network Control Protocol* (NPC) (CARR; CROCKER; CERF, 1970). O NPC tornou possível a implementação das aplicações distribuídas que surgiram nos dois anos seguintes na ARPA Net.

A ARPA Net foi uma rede experimental, precursora da Internet, usada por alguns cientistas da computação e membros do Departamento de Defesa (DoD) estadunidense. Nela se consolidaram as primeiras tecnologias de rede baseadas em roteamento de pacotes.

A primeira demonstração pública das capacidades da ARPA Net foi realizada por Robert Kahn na *I*nternational Computer Communication Conference (ICCC), no ano de

1972. No mesmo ano em que Ray Tomlinson (BBN Technologies), com a intenção de obter um mecanismo para comunicação entre os desenvolvedores da ARPA Net, elaborou e tornou público o primeiro programa para envio e recebimento de emails (LEINER et al., 2009).

Com o passar do tempo, a ARPA Net original evoluiu para a Internet, cuja principal diferença em relação a sua antecessora foi idealizar um sistema em que redes de diferentes padrões pudessem se conectar.

2.1.3 Idealização

Para a Internet, Kahn desenvolveu um novo protocolo de comutação de pacotes chamado *Transmission Control Protocol/Internet Protocol* (TCP/IP) que se baseava em quatro aspectos fundamentais (CERF; KAHN, 1974):

- Evitar que mudanças internas fossem requeridas nas diferentes redes que se integrariam à Internet, mantendo a independência de funcionamento de cada uma dessas redes;
- Transmitir dados baseados em comutação de pacotes, com necessidade eventual de retransmissão em caso de perda do pacote;
- Interconectar diversas redes através de caixas-pretas que teriam como função redirecionar os pacotes e recuperar a comunicação em caso de falhas, como a perda de pacotes. Posteriormente, essas ideias resultaram no desenvolvimento de dispositivos físicos e virtuais de rede como *gateways, switches* e *routers*, cada um deles com suas particularidades;
- Adotar uma infraestrutura de comunicação com governança decentralizada.

O conjunto de protocolos TCP/IP também endereçava:

- Transmissão de pacotes entre roteadores conectados indiretamente, através de outros roteadores, que funcionam nesse caso como pontes entre os pontos de comunicação. Fora o redirecionamento dos pacotes ao longo de possivelmente múltiplos *hops*, a interconexão exigia o tratamento de aspectos como fragmentação de pacotes;
- Algoritmos para detecção e retransmissão de pacotes perdidos;
- Técnicas para controle de fluxo e detecção de congestionamento que permitem que os pacotes sejam escoados através de caminhos otimizados;
- Sistema global de endereçamento;

- Esquemas de detecção e tratamento de erros ponto-a-ponto, que exerceram um papel fundamental em um contexto onde os dispositivos físicos possuiam estabilidade e confiabilidade muito inferiores ao que se possui na atualidade;

- Secundariamente, aspectos como eficiência e segurança também foram considerados, mas com menor importância.

A colaboração entre Kahn e Vint Cerf teve inicio em 1973. Cerf possuia conhecimentos sólidos sobre sistemas operacionais e havia participado do projeto e desenvolvimento do *Network Control Protocol* (NCP), o antecessor do TCP. Como resultado, o protocolo adotou alguns princípios, como:

- A comunicação entre dois processos se daria por meio de uma stream de bytes (octetos);

- O controle de fluxo seria realizado através de uma janela deslizante que marca o último pacote com confirmação de recebimento e o último pacote enviado. Confirmações de recebimento deveriam ser enviadas, cada confirmação significando que todos os pacotes até o confirmado foram recebidos com sucesso;

- Os parâmetros da janela deslizante não seriam definidos, mas deixados a critério dos *endpoits*. As diversas abordagens implementadas para realizar controle de fluxo deram orígem às diferentes variantes do protocolo TCP/IP, como: Tahoe, Reno, New Reno, Sack, Westwood, Veno, entre outras (Kassem et al., 2010).

- Para cada nó da rede, seria atribuído um endereço IP de 32 bits, onde os 8 bits mais significativos caracterizavam a rede e os 24 bits restantes identificavam o host dentro daquela rede. O projeto inicial para o IP considerava o contexto de uma rede nacional da época, como a ARPA Net, de modo que 256 subredes parecia ser um número razoável. Verdade que mudou no fim dos anos 70, quando se popularizou o protocolo Ethernet, fazendo crescer drasticamente o número de Redes de Acesso Local (LANs).

Anos depois, a proliferação de sistemas que faziam uso da pilha TCP/IP evidenciou pontos positivos e negativos do TCP. O protocolo mostrou-se adequado para aplicações de transferência de arquivos e acesso à estações remotas, por exemplo. Entretanto, para casos como o dos sistemas de comunicação por voz em tempo real, ficou claro que o tratamento da perda de pacotes deveria ser gerenciada pela camada de aplicação. Devido a isso, criou-se um protocolo de transferência alternativo para prover acesso à camada de IP. Surgia o *User Datagram Protocol* (UDP) que diferenciou-se do TCP, principalmente pela característica ausência da confirmação de recebimento de pacotes.

2.1.4 Implementação

Os resultados da colaboração entre Cerf e Kahn foram publicados em (CERF; KAHN, 1974) e, apenas um ano depois, já haviam três implementações distintas e interoperantes.

Começando com apenas três redes (ARPA Net, Packet Radio e Packet Satellite), a Internet foi rapidamente incorporando outras redes da época. Uma delas foi a Ethernet, criada em 1973 por Bob Metcalf na Xerox PARC. A inclusão dessa rede e o barateamento dos computadores permitiram que se desenvolvessem um grande número de LANs, aumentando drasticamente o número de pontos de acesso à Internet. Para acomodar o grande número de redes que surgiam, dividiu-se as redes em três classes:

- **Classe A:** redes de escala nacional (2^7=128 redes com 2^{24}=16.777.216 endereços por rede);
- **Classe B:** redes de escala regional (2^{14}=16.384 redes com 2^{26} endereços por rede);
- **Classe C:** redes de acesso local (2^{21}=2.097.152 redes com 2^8=256 endereços por rede).

Visando tornar o acesso à Internet mais amigável ao uso humano, atribuiu-se nomes aos hosts. Na ARPA Net, esse mapeamento era descrito em um único *host*. Com o crescimento do número de nós conectados à Internet, essa abordagem tornou-se inviável.

Em 1983, Paul Mockapetris apresentou o *Domain Name System* (DNS) para resolução desse problema. Esse sistema consistia em uma cadeia distribuída e escalável de servidores que permitiam a resolução de nomes de forma hierárquica.

2.1.5 Consolidação

Em 1985, a Internet já estava consolidada como tecnologia. Nos anos que se seguiram, as infraestruturas dedicadas foram se propagando de forma heterogênea, de acordo com a região geopolítica. Nos Estados Unidos, em particular, tem-se como marco histórico a criação da NSFNET em 1986. A princípio, essa rede interligava cinco grandes centros computacionais com uma banda de 56 kilobits por segundo. No começo dos anos 90, a mesma banda havia sido aumentada para 45 megabits por segundo. Em 1995 a NSF transferiu a NSFNET para o setor comercial, que se tornou o principal *backbone* da Internet (FOSTER; KESSELMAN, 2003).

Para ampliar o alcançe da Internet, no início, companhias de telecomunicações utilizaram a mesma infraestrutura da *Public Switched Telephone Network* (PSTN) para prover serviços de conexão à Internet através da tecnologia de acesso Dial-up. Após a virada do milênio, devido a limitação de velocidade imposta pelas linhas de cobre, popularizaram-se nos países mais desenvolvidos as conexões de banda-larga que garantiam taxas de até 512 kilobits por segundo (MURRAY-WEST, 2016).

Do ponto de vista de usabilidade, a Internet teve um avanço significativo quando, em 1989, Tim-Berners Lee, trabalhando no projeto CERN, propôs a WWW(BERNERS-LEE, 1989). O propósito original era facilitar a comunicação entre pesquisadores e engenheiros do projeto, através da criação de um repositório global de conhecimento. Para isso, foram criados diversos padrões e protocolos. Alguns deles, como *Universal Resource Identifier* (URI), *Hypertext Transfer Protocol* (HTTP) e *HyperText Markup Language* (HTML), persistem até os dias atuais, com atualizações. Algumas das principais características da WWW são:

- O sistema de endereçamento universal URI, que torna todos os recursos do repositório enderessáveis[1];
- O protocolo de comunicação HTTP, utilizado para padronizar as requisições e respostas entre os clientes e servidores;
- A linguagem de marcação HTML, que fornece um meio para descrever páginas e links, com a utilização de tags descritivas de *layout* que devem ser renderizadas no lado do cliente através do uso de *browsers*(BERNERS-LEE et al., 1994).

2.1.6 Contemporaneidade

O estabelecimento de transmissores de comunicação sem fio na superfície da Terra e em satélites artificiais permitiu que a conexão de Internet ganhasse disponibilidade global. O aprimoramento das tecnologias associadas à fibra ótica e produção de circuitos integrados permitiu que o acesso à Internet chegasse à taxas de transmissão inimagináveis no início (como o recorde mundial atual de 178 Terabits por segundo (UCL, 2020)).

Atualmente, a Web é o maior repositório de informação que já existiu em toda história da humanidade. Uma estimativa recente indica o número de 4.66 bilhões de usuários ativos, dos quais 4.32 bilhões navegam em telefones móveis e 4.2 bilhões fazem uso interativo da Web. (STATISTA, 2021a).

2.2 COMPUTAÇÃO EM GRADE

A ideia de federar recursos computacionais de diferentes características e organizações levou à concepção da computação em grade, no ano de 1998. Nessa seção, apresenta-se uma definição e um histórico dessa tecnologia segundo as refêrencias providas pelos pesquisadores pioneiros da área (FOSTER, 2002) (FOSTER; KESSELMAN, 2003) (FOSTER; KESSELMAN, 2011).

[1] Uma URL é um caso particular de utilização de uma URI, popularizado principalmente pelo seu uso em protocolos de comunicação fundamentais como o HTTP e FTP.

2.2.1 Conceituação

Em (FOSTER, 2002), Ian Foster define a computação em grade por meio de três características principais:

1. **Coordenação de recursos que não estão sujeitos a um controle centralizado ...** Uma grade computacional integra e coordena recursos e usuários situados em domínios de controle distintos.

2. **... usando protocolos e interfaces padronizados, abertos e de propósito geral ...** Um Grid é construído usando interfaces e protocolos multiuso que endereçam problemas fundamentais como autenticação, autorização, descoberta e acesso a recursos computacionais. Foster entende que é importante que as interfaces e protocolos sejam padronizados e abertos, caso contrário estaríamos lidando com um sistema de aplicação específica.

3. **... para entregar qualidade não-trivial de serviços.** Um Grid permite que seus recursos constituintes sejam usados de modo coordenado para entregar alta qualidade de serviço considerando, por exemplo, tempo de resposta, *throughput*, disponibilidade e segurança.

Tipicamente, as grades estão dispostas em quatro classes (FOSTER; KESSELMAN, 2003):

- **Grade computacional:** uma infraestrutura de gerenciamento de recursos distribuídos que tem como foco coordenar o acesso a recursos computacionais remotos;

- **Grade de desktops:** grades computacionais que mapeiam coleções de tarefas desacopladas para recursos não-dedicados, como um desktop pessoal;

- **Grade de dados:** grade composta por serviços de gerenciamento de dados federados;

- **Grade de serviços:** uma infraestrutura que tem como objetivo confederar coleções de Web Services de aplicações específicas.

A intersecção da computação em grade e a computação na nuvem se dá principalmente no contexto corporativo, onde as "Grades Corporativas" tranformaram-se em "Nuvens Privadas" após agregarem a virtualização para facilitar o provisionamento dinâmico de recursos (FOSTER; KESSELMAN, 2011).

2.2.2 Síntese Histórica

Desde o fim da década de 80, pesquisadores visionam uma infraestrutura computacional que fornecesse acesso à computação sob demanda e permitisse um compartilhamento

de recursos de modo flexível, seguro e coordenado entre conjuntos dinâmicos de indivíduos, instituições e recursos. Essa visão foi chamada de "grade" em referência às redes (ou grades) de distribuição de energia elétrica.

Uma das primeiras tentativas de aproveitar-se de diversos computadores conectados em rede para execução de uma tarefa se deu com a implementação das plataformas HTCondor (originalmente chamada Condor), em 1988, e Load Sharing Facility (originalmente chamada de Utopia (ZHOU et al., 1993)).

Com a popularização da Web na década de 90, uma grande gama de projetos surgiram para se aproveitar dos recursos dela para realizar computação remota, por exemplo: Charlotte (KEDEM; WYCKOFF, 1996)(BARATLOO et al., 1999), ParaWeb (BRECHT et al., 1996), Popcorn (CAMIEL, 1997) e SuperWeb (ALEXANDROV et al., 1997). O sugimento de redes de alta velocidade, como o banco de ensaios AURORA gigabit (CLARK et al., 1993), tornou possível a interconexão de múltiplos *sites* para compartilhamento de recursos computacionais de processamento.

Um ambiente experimental de computação em grade chamado I-WAY foi criado em 1994 (FOSTER; KESSELMAN, 2003). Seu objetivo era permitir que pesquisadores utilizassem múltiplos supercomputadores interconectados para processar algoritmos de visão computacional. O I-WAY integrou 12 bancos de ensaios do Gerenciamento de Tráfego Aéreo Estadunidense (ATM), 17 centros com supercomputadores, 5 *sites* para pesquisa de Realidade Virtual (VR) e mais de 60 grupos de aplicações. Além de bem sucedido, o experimento incentivou o desenvolvimento da infraestrutura I-Soft, uma precursora do Globus Toolkit e da National Technology Grid.

O Globus Toolkit (GT) consistia em um conjunto de ferramentas de baixo-nível contendo facilidades e mecanismos básicos para o desenvolvimento de aplicações executadas em grade, como: comunicação, autenticação, informações de rede e acesso a dados remotos (FOSTER; KESSELMAN, 1997). Em sua quarta versão, o GT enderessava em seu *core* questões como: segurança; descoberta, gerenciamento e acesso a recursos (FOSTER, 2006).

Nesse interim, comunidades científicas visavam a computação em grade como solução para o problema da federação de recursos. Pesquisadores criaram na Europa o projeto EU DataGrid (GAGLIARDI et al., 2002) com objetivo de realizar as computações referentes aos experimentos do Large Hadron Collider (LHC). Nos Estados Unidos, a união das organizações Particle Physics Data Grid e Grid Physics Network levaram à criação do Open Science Grid (AVERY et al., 2001).

Uma das primeiras iniciativas de computação em grade voluntária se deu através do projeto SETI@home. A ideia era elaborar um sistema onde milhões de unidades computacionais, com características domésticas e heterogêneas, pudessem contribuir com tarefas de computação que analisavam sinais eletromagnéticos do espaço (ANDERSON et al., 2002).

Com o propósito de oferecer suporte ao projeto SETI@home, e muitos outros projetos de computação em grade volutária que surgiriam nos anos seguintes, a U.C. Berkeley, criou em 2002 uma plataforma chamada de Berkeley Open Infrastructure for Network Computing (BOINC). A plataforma simplificou os processos de criação e operação de projetos de computação em grade voluntária. Fazendo uso de clientes locais, os proprietários de máquinas pessoais podem contribuir com poder de processamento em diversos projetos BOINC e especificar o provisionamento de recursos locais de acordo com a finalidade da pesquisa (Anderson, 2004).

2.2.3 Contemporaneidade

A plataforma BOINC continua em operação até os dias atuais. Atualmente, existem mais de 30 projetos de pesquisa ativos nas mais diversas áreas do conhecimento científico reconhecidos pela U.C. Berkley. As áreas incluem: Física, Astrofísica, Matemática, Ciência da Computação. Inteligência Artificial, Biologia, Medicina, entre outras (BERKELEY, 2021).

Um dos projetos BOINC mais bem sucedidos é o World Community Grid (WCG), dirigido pela IBM. Atualmente, ele conta com a contribuição de mais de 650 mil indivíduos e 460 organizações. Até o momento, o WCG dá suporte a 31 projetos. Entre as iniciativas atuais, incluem-se pesquisas sobre o Câncer, Microbiologia e a COVID-19 (WCG, 2021). A Figura 1 ilustra a facilidade de que dispõem os proprietários de computadores pessoias para contribuir em projetos da plataforma BOINC.

2.3 VIRTUALIZAÇÃO

A última das tecnologias fundamentais avaliadas nesse capítulo é a virtualização. Por meio do uso de técnicas de virtualização, múltiplos inquilinos podem utilizar simultaneamente os mesmos recursos físicos computacionais, em contextos completamente isolados. A virtualização também permite que recursos sejam provisionados de forma automatizada e programática, promovendo o desacoplamento entre a infraestrutura da nuvem e os dispositivos físicos que a compõem.

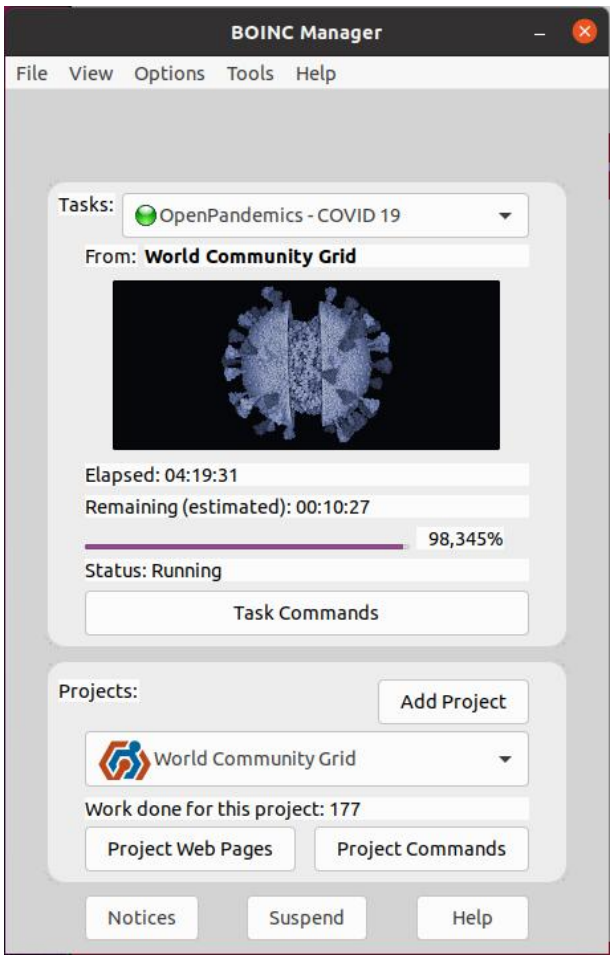
Por definição, a infraestrutura de nuvem é composta por uma *pool* de recursos virtualizados (ver Capítulo 3). A implementação dessa *pool* só é possível por meio do uso de técnicas de virtualização.

Nessa seção, apresenta-se uma definição e um breve resumo histórico das tecnologias de virtualização.

2.3.1 Conceituação

Em 2006, Campbell e Jeronimo definem a virtualização precisa e sucintamente como "o processo de desacoplamento entre hardware e sistema operacional em uma máquina física"

Figura 1 – BOINC Manager: um cliente moderno para contribuir com projetos BOINC através de uma interface gráfica. O screenshot foi gerado em uma máquina pessoal com o sistema operacional Ubuntu 20.04 LTS e o cliente BOINC Manager na versão 7.16.6. Na imagem, a máquina trabalha em tarefas do projeto OpenPandemics - COVID 19, que tem como propósito analisar dados de simulações moleculares em busca de possíveis tratamentos para a doença.



Fonte: (WORLDCOMMUNITYGRID, 2021)

.

(CAMPBELL; JERONIMO, 2006). Desde então, muitas outras formas de virtualização surgiram. Por essa razão, esse trabalho utiliza uma definição mais ampla de virtualização:

Definição 2 *Virtualização é o processo de desacoplamento entre os recursos físicos e os recursos lógicos oferecidos por um sistema computacional.*

Técnicas de virtualização podem ser empregadas para criar abstrações sobre recursos como memória, redes, armazenamento, hardware, sistemas operacionais, aplicações e ambientes de execução (SAREEN, 2013).

Uma possível classificação para os tipos existentes de virtualização é dada por (REDHAT, 2021d):

- **Virtualização de Dados:** Dados espalhados lógica e geograficamente podem ser consolidados em uma única entidade virtual;
- **Virtualização de *Desktop*:** Ambientes virtuais de trabalho desacoplados de dispositivos físicos;
- **Virtualização de Servidor:** A capacidade de computação dos servidores pode ser usada para atender simultaneamente a demanda computacional exigida por aplicações distintas;
- **Virtualização de Sistemas Operacionais:** As funções do *kernel* de um sistema operacional rodam sobre uma infraestrutura virtual, permitindo a criação de ambientes com alto nível de isolamento com o *host*. Quando esses ambientes isolados compartilham o código do *kernel* da máquina host, dá-se o nome de contêiner ao ambiente virtualizado (ver Subseção 2.3.2);
- **Virtualização de Funções de Rede (NFV):** Funções de rede como serviços de diretório, compatilhamento de arquivos e configuração de IPs podem ser distribuídas programaticamente em ambientes distintos.

A utilidade da virtualização pode ser observada através de diversas aplicações. Por exemplo (SUSANTA; TZI-CKER, 2005):

- Consolidação de servidores e aplicações, para otimização do aproveiramento da infraestrutura;
- Implementação de *sandboxes* (ambientes de execução efêmeros para testes) e aplicações *multitenant* (a mesma instância de aplicação serve diversos inquilinos, cada um com configurações e estados próprios);
- Gestão centralizada de múltiplos ambientes de execução, como os ambientes de desenvolvimento, *staging* e produção;

- Virtualização de diversos dispositivos de hardware para facilitar a depuração de componentes de baixo-nível;
- Execução simultânea de sistemas operacionais distintos no mesmo *desktop*;
- Conveniências na execução de migrações de software, como *rollback* (retornar o sistema para um estado anterior no tempo);
- Facilidades na criação de cenários de teste, execução de *replays* dos casos analisados e depuração erros.

Ao longo da evolução das técnicas de virtualização de máquinas, criou-se uma terminologia especial. Alguns dos conceitos mais fundamentais são (CAMPBELL; JERONIMO, 2006):

- **Máquina *Host*:** É um máquina física com recursos físicos, como memória, armazenamento, interface de rede e CPU, sobre a qual roda-se o software de virtualização;
- **Máquina Virtual:** A máquina obtida a partir da abstração provida pelo *software* de virtualização rodando sobre uma máquina *host*;
- **Sistema Operacional *Guest*:** sistema operacional utilizado diretamente pela máquina virtual;
- ***Software* de Virtualização:** Termo genérico para denominar softwares que permitem rodar máquinas virtuais sobre a máquina host;
- **Disco Virtual:** Abstração utilizada em softwares de virtualização para permitir que uma parte do armazenamento da máquina host se comporte como uma unidade de armazenamento íntegral para a máquina virtual. Esse recurso facilita a portabilidade e replicação em lote desses ambientes.
- ***Additions* da Máquina Virtual:** Conjunto de recursos para melhorar a eficiência e integração do sistema operacional da máquina host com o software de virtualização;
- **Diretórios compartilhados:** Geralmente disponível após a instalação das Additions, o recurso de compartilhamento de diretórios permite que diretórios da máquina host sejam mapeados em diretórios da máquina virtual, simplificando o compartilhamento de arquivos entre elas;
- **Monitor de Máquina Virtual (VMM) ou *Hosted Hypervisor*:** Software de virtualização de hardware que permite executar e gerenciar máquinas virtuais rodando sobre um sistema operacional *host* (ex.: Oracle VM VirtualBox);

- ***Hypervisor* ou *Bare-Metal Hypervisor*:** Software de virtualização que executa diretamente sobre o hardware, sem interação ou concorrência com um sistema operacional na máquina host (ex.: VMware ESXi);

- **Paravirtualização:** Conjunto de modificações realizadas na imagem de um sistema operacional para otimizar sua instalação e funcionamento em uma máquina virtual. As modificações incluem, em grande parte, a inclusão de rotinas para melhorar a eficiência da comunicação do sistema operacional *guest*, com a plataforma de virtualização utilizada;

Embora não seja um conceito exclusivo de máquinas virtuais ou containeres, uma propriedade extremamente desejável da perspectiva de segurança é que haja isolamento entre os ambientes dos múltiplos inquilinos instanciados sobre a mesma infraestrutura física de computação.

2.3.2 Máquinas Virtuais e Contêineres

Máquinas virtuais são ambientes virtualizados que incluem cópias completas dos sistemas operacionais, conhecidas como *Guest OS*. Em decorrência disso, o tamanho típico de armazenamento requerido para máquinas virtuais é da ordem de GBs [2]. Contêineres são uma abstração operando sobre o sistema operacional para fornecer um ambiente homogêneo para desenvolvimento, teste e operação de aplicações.

Múltiplos contêineres podem rodar simultaneamente na mesma máquina física, compartilhando a mesma instância de Kernel com outros contêineres. Essa abordagem reduz drasticamenticamente o tamanho das imagens para a ordem de MBs[3] e otimiza a execução concorrente através do compartilhamento das funções comuns de *kernel* do sistema operacional do *Host*. A Figura 2 explicita as diferenças entre os modelos de virtualização analisados, apresentando as camadas de software envolvidas na composição de cada um dos ambientes virtuais. No exemplo, a plataforma de conteinerização descrita é Docker.

2.3.3 Síntese Histórica

De acordo com (CAMPBELL; JERONIMO, 2006), a história da virtualização inicia-se no final dos anos 50 quando um grupo da Universidade de Manchester desenvolveu um sistema para substituição automatizada de páginas de memória para o *mainframe* Atlas, inaugurando o conceito de memória virtual.

Quase 10 anos depois, em 1967, a IBM apresentou o *mainframe* System/360 modelo 67, a primeira versão *major* do sistema que contava com memória virtual. Uma inovação

[2] A imagem da versão padrão do Ubuntu Desktop 20.04 é cerca de 2.7 GB (CANNONICAL, 2021)

[3] A execução do comando "docker images" retorna o valor de 72.7 MB para tamanho da imagem Docker (cerca de 2.6% do tamanho das imagens do sistema operacional) para a imagem 'ubuntu', em sua versão mais recente "7e0aa2d69a15".

desse modelo foi operar sobre um conjunto de instruções virtualizadas através do uso do sistema operacional CP-67 que evoluiu para uma família de sistemas operacionais chamada "VM". Esses sistemas permitiram a execução concorrente de instâncias isoladas de sistemas operacionais, possivelmente distintos, no mesmo *mainframe*. Na mesma década, emergiu o conceito de virtualização de hardware que permitiu ao monitor de máquinas virtuais rodar instâncias em ambientes protegidos e isolados.

Em meados dos anos 70, a virtualização já era uma prática bem aceita por usuários de mainframes de diversos sistemas operacionais. Em 1979, o utilitário *chroot* foi introduzido ao UNIX versão 7. Por meio do uso desse utilitário, administradores UNIX estabeleceram um meio para controlar a segurança de um software. Para isso, cada aplicação era colocada em uma *jail* distinta que limitava a visão da aplicação em relação ao sistema de arquivos. Esse utilitário foi nesse sentido um precursor das técnicas de conteinerização modernas.

Com o barateamento da tecnologia, a virtualização saiu de foco nos anos 80, a medida em que os computadores pessoais surgiram e se tornavam um dos principais bens de consumo da economia mundial.

O uso de virtualização foi retomado durante os anos 90. Uma das principais motivações era rodar aplicações legadas em sistemas operacionais mais recentes. Por meio da virtualização, foi aumentada a flexibilidade dos servidores para rodar sistemas de diferentes fornecedores e utilizar de forma mais eficiente a infraestrutura de *hardware*, que agora podia ser compartilhada.

Na segunda metade dos anos 90, emergiu uma nova categoria de máquinas virtuais. Diferente das aplicações convencionais da época, essas máquinas não apresentavam uma plataforma virtual de hardware. Ao invés disso, faziam uso dos benefícios de homogeneidade providos pelas máquinas virtuais para criar ambientes homogêneos de execução de desenvolvimento e execução de software. Dentre as plataformas mais populares, destacam-se a Java Virtual Machine (JVM), introduzida em 2004, e a Microsoft Common Language Runtime (CLR) para ambientes Windows.

2.3.4 Contemporaneidade

O interesse por técnicas de virtualização de sistemas operacionais foi renovado no fim dos anos 90, resultando principalmente na criação de diversas plataformas para virtualização de máquinas como a VMware Workstation, em 1998 (VMWARE, 2021), e, posteriormente, as plataformas VirtualBox (WIKIPEDIA, 2021h) e Hyper-V (MICROSOFT, 2008), em 2008, mesmo ano em que foi lançada a plataforma Linux Containers (LXC) (WIKIPEDIA, 2021f).

A LXC realiza virtualização a nível de sistema operacional através da criação de instâncias de contêineres definidos a partir de imagens pré-existentes. O uso de contêineres mostrou-se apropriado para o empacotamento de alguns tipos de aplicações. Em espe-

cial os sistemas arquitetados sob o paradigma de microsserviços. Desde então surgiram diversas plataformas para conteinerização.

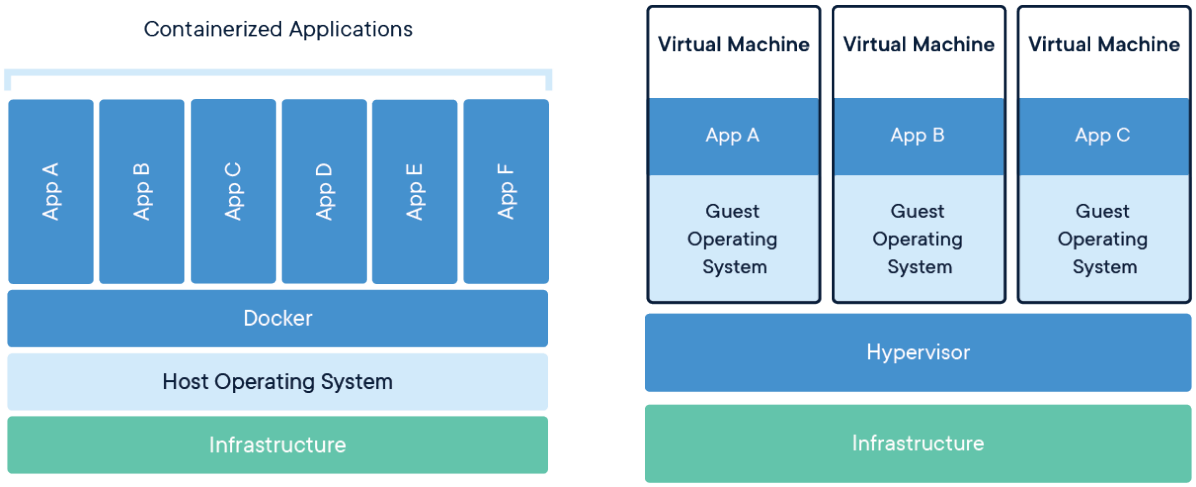
Lançada em 2011, a plataforma mais popular para conteinerização de aplicações na atualidade é o Docker. O elevado número de contêineres existentes em sistemas de grande porte, principalmente em decorrência das arquiteturas de microsserviços, tornou necessária a criação de ferramentas mais complexas para a gerência de *deployment* (WIKIPEDIA, 2021c).

Lançado em 2014, o Kubernetes (k8s) é um sitema open-source para automação de *deployment*, escalabilidade e gerenciamento de aplicações conteinerizadas (KUBERNETES, 2014). Atualmente, o k8s é uma das plataformas mais populares para orquestração de contêineres. Alguns dos recursos da plataforma são (GURU99, 2021):

- Desacoplamento entre aplicação e a infraestrutura;
- Escalonamento automático das unidades de deployment (pods) nas unidades de computação (nós);
- Monitoramento e regeneração automatizada de serviços com falha;
- *Rollback* e *rollout* manual e automático de *deployment*;
- Balanceamento de carga embutido como componente nativo da plataforma;
- Escalabilidade horizontal automatizada;
- Consitência entre ambientes de desenvolvimento, teste e produção.

O conceito *Infrastructure as Code* (IaC) foi criado para descrever plataformas de orquestração que permitem a definição de uma infraestrutura virtual através de uma linguagem de programação. Para esse propósito, algumas das ferramentas mais populares na atualidade são: Ansible, Puppet, Salt, Terraform e AWS CloudFormation (WIKIPEDIA, 2021e).

Figura 2 – Representação gráfica que ilustra a diferença entre as camadas de execução em esquemas de virtualização baseados em contêineres (à esquerda), e em máquinas virtuais (à direita).



Fonte: (DOCKER, 2021)

.

3 CONCEITUAÇÃO

Nesse capítulo, conceitua-se formalmente o modelo de computação na nuvem por meio da descrição de cada uma das suas propriedades fundamentais. Adicionalmente, exemplos são citados ao longo do capítulo para expandir e contextualizar a compreensão do leitor.

Os aforismos entre aspas desse capítulo que definem as características essenciais da nuvem, descendem diretamente de uma tradução livre de (MELL; GRANCE et al., 2011).

3.1 DEFINIÇÃO GERAL

Em 2011, o NIST definiu o modelo de computação na nuvem formalmente como (MELL; GRANCE et al., 2011):

Definição 3 *"**Computação na nuvem** é um modelo para tornar uma pool compartilhada de recursos computacionais (como redes, servidores, armazenamento, aplicações e serviços) acessível através por meio de uma rede de computadores, de forma ubíqua, conveniente e sob-demanda. O provisionamento e liberação desses recursos deve ocorrer de forma rápida, com o mínimo de esforços de gerenciamento e de interação com o provedor de serviços. Esse modelo de nuvem é composto por cinco características essenciais, três modelos de serviço e quatro modelos de deployment."*

3.2 AS CINCO CARACTERÍSTICAS ESSENCIAIS DA NUVEM

Essa seção apresenta as cinco características essenciais da nuvem, de acordo com o modelo proposto pelo NIST em 2011.

3.2.1 *Self-service* sob-demanda

"**Um consumidor pode provisionar unilateralmente capacidades computacionais (como tempo em servidor ou armazenamento em rede), de acordo com sua necessidade sem requerer interação humana com os provedores de serviços.**" (MELL; GRANCE et al., 2011)

Significa que todo provisionamento de recursos e uso dos serviços na nuvem devem ocorrer sem necessidade de comunicação entre o consumidor com um agente humano do provedor, permitindo dessa forma a elevação do grau de automação nos processos de desenvolvimento, *deployment* e operação dos sistemas desse modelo.

A gerência do provisionamento pode ocorrer de forma automática (ver Subseção 3.2.4) ou manual (manipulando interfaces gráficas de usuário ou interfaces programáticas). As interfaces de governança e gerenciamento manual oferecidas pelos provedores públicos geralmente são:

- **Interface gráfica de usuário (GUI):** aplicações web como AWS Console (AWS, 2021i), Azure Portal (AZURE, 2021c) e GCP Console (GCP, 2021h);
- **Interface de linha de comando (CLI):** AWS CLI (AWS, 2021h), Azure CLI (AZURE, 2021b) e GCP CLI (GCP, 2021f);
- **Interface de programação de aplicação (API):** AWS APIs (AWS, 2021g), Azure APIs (AZURE, 2021a) e GCP APIs (GCP, 2021g).

3.2.2 Amplo acesso à rede

"**Capacidades computacionais disponíveis através de uma rede de computadpres e acessadas através de mecanismos padrões que permitem o uso em plataformas heterogêneas do cliente.**"(MELL; GRANCE et al., 2011)

Para alcançar a ubiquidade na provisão dos serviços, o modelo de computação na nuvem se apoia fortemente na premissa de que ambas partes (provedor e consumidor) possuem acesso estável e de alta velocidade a uma rede em comum.

No caso do modelo de *deployment* público, essa rede é, frequentemente, a Internet (LIU et al., 2011). Por essa razão, o modelo público de *deployment* só conseguiu se popularizar a partir dos anos 2000 (ver Seções 2.1 e 4.3).

Uma estratégia utilizada por alguns provedores públicos para garantir melhor latência, estabilidade e desempenho de rede em geral, é possuir infraestrutura privada, geograficamente próxima dos clientes (ver Seção 5.5).

3.2.3 *Pooling* de recursos

"**Os recursos computacionais do provedor estão dispostos em *pools*[1] para servir múltiplos consumidores usando um modelo *multi-tenant*, com diferentes recursos físicos e virtuais sendo alocados e desalocados dinamicamente de acordo com a demanda do consumidor. Aqui há a ideia de independência de localização do recurso físico, na qual o consumidor geralmente não tem controle ou conhecimento sobre a localização exata dos recursos providos. Em alguns casos, é possível especificar uma localização em alto-nível (como país, estado ou *datacenter*). Exemplos de recursos incluem armazenamento, processamento, memória e banda de rede.**"(MELL; GRANCE et al., 2011)

Essa é um propriedade que viabiliza a computação na nuvem do ponto de vista econômico. Fazendo uso de uma *pool* de recursos virtualizados, os recursos físicos podem ser

[1] No contexto de ciência da computação, o termo *pool* refere-se a uma coleção de recursos que é mantida em estado de prontidão para uso. Quando um cliente requisita um recurso virtual, ele é alocado a partir da *pool* (ao invés de instanciar um novo recurso) e, após o uso, devolvido para a mesma *pool* (ao invés de ser destruído), para poder ser utilizado por outros clientes (WIKIPEDIA, 2021g).

compartilhados e utilizados por diversos consumidores simultaneamente, de modo a otimizar a utilização. Como resultado, nota-se que, no modelo de computação na nuvem, muito mais clientes podem ser atendidos, se comparado com o modelo de provisionamento dedicado.

A tecnologia que torna esse compartilhamento de recursos físicos possível é a virtualização. Através da virtualização, tem-se o desacoplamento entre as camadas de serviço e hardware do provedor e, desse modo, permite-se que uma unidade física do recurso hospede diversas unidades virtuais (ver Seção 2.3.1).

3.2.4 Rápida elasticidade

"**Capacidades computacionais podem ser provisionadas de forma elástica e liberadas, em muitos casos automaticamente, para escalar rapidamente e de modo proporcional à demanda. Para o consumidor, as capacidades disponíveis para provisionamento parecem ser ilimitadas e podem ser apropriadas em qualquer quantidade e durante qualquer período de tempo.**"(MELL; GRANCE et al., 2011)

Trata-se da propriedade da computação na nuvem que torna os serviços altamente escaláveis, nesse modelo de computação. Significa que a alocação de recursos computacionais pode ser ajustada para acompanhar as mudanças na demanda do serviço. Usualmente, essa propriedade é alcançada com o uso de ferramentas de orquestração de aplicações e *deployment* automatizado.

Dessa maneira, o sistema do consumidor pode fazer uso de recursos computacionais intensamente durante os períodos de pico de utilização, e liberá-los de volta para a *pool* de recursos, após a diminuição do volume de utilização. De modo que, o consumidor pode obter um valor menor de tarifação, em relação à alocação exclusiva de recursos computacionais (a depender do perfil de utilização do consumidor) (ver Seção 5.6).

É comum que os provedores de computação na nuvem ofereçam ferramentas e serviços que permitem que o gerenciamento da escalabilidade seja realizado de forma automatizada e, em muitos casos, até mesmo assistida por aprendizado de máquina (escalabilidade preditiva) (AWS, 2021q).

3.2.5 Serviços com métricas

"**Sistemas na nuvem automaticamente controlam e otimizam o uso de recursos adicionando capacidades de mensuração em algum nível de abstração apropriado ao tipo de serviço. O uso de recursos pode ser monitorado, controlado e reportado, provendo transparência tanto para o provedor quanto para o consumidor do serviço em utilização.**"(MELL; GRANCE et al., 2011)

Métricas de serviços são coleções de dados, referentes ao funcionamento e desempenho dos recursos monitorados. Geralmente, essas métricas são organizadas como séries temporais que descrevem a variação de alguma propriedade do recurso ao longo do tempo (GCP, 2021i).

As métricas de serviço desempenham um papel importante na computação na nuvem. Por meio da leitura dessas métricas, é possível:

- Implementar o modelo de cobrança *pay-per-use*, onde o consumidor é cobrado apenas pelas janelas de tempo em que fez uso. Atualmente, provedores públicos oferecem janelas de até 1 segundo de precisão para tarifação (AWS, 2021j);
- Realizar um monitoramento contínuo do estado de disponibilidade e saúde de cada um dos serviços, em cada uma das regiões e zonas de disponibilidade em que eles são oferecidos (ver Seção 5.3);
- Permitir que os provedores dêem respostas rápidas aos incidentes gerados por indisponibilidade ou degradação dos serviços (ver Subseção 5.4.2);
- Reconhecer casos onde os acordos de nível de serviço (SLAs) foram violados (ver Subseção 5.4.1);
- Identificar problemas de desempenho de rede nos serviços providos (ver Subseção 5.5).

3.3 MODELOS DE SERVIÇO

Os modelos de serviço da computação na nuvem caracterizam a camada de interação entre o provedor e o consumidor de serviços, bem como o nível de abstração provido pelo serviço.

O modelo de computação na nuvem distingue quatro camadas de tecnologias(TANENBAUM; STEEN, 2007):

- **Hardware:** Geralmente instalada em grandes *datacenters*, a camada de Hardware é composta por dispositivos físicos como servidores, roteadores, fontes elétricas e sistemas de refrigeração. No modelo de computação na nuvem, essa camada não está diretamente acessível ao consumidor;
- **Infraestrutura:** Empregando técnicas de virtualização sobre a camada de hardware, obtem-se a camada de Infraestrutura, que pode ser oferecida diretamente como serviço (ver Subseção 3.3.1);
- **Plataforma:** Ambiente virtualizado sobre a camada de Infraestrutura da nuvem. É constituído de ferramentas para desenvolvimento e *deploy* de aplicações na nuvem,

bem como recursos de armazenamento com alto-nível de abstração como *storage buckets* e bases de dados.

- **Aplicação:** Instâncias das aplicações rodando sobre a camada de Plataforma. Aplicações dessa camada têm todo seu processamento realizado no provedor e a máquina do cliente é apenas utilizada como um terminal de interface.

3.3.1 Infrastrutura como um Serviço (IaaS)

"**Serviço que provê ao consumidor capacidade de provisionamento de recursos como processamento, armazenamento e rede. Através do uso desses recursos, o consumidor pode fazer *deploy* e rodar softwares arbitrários como sistemas operacionais e aplicações. Embora o consumidor não gerencie ou controle a infraestrutura subjacente da nuvem (como os dispositivos físicos), ele tem controle sobre o sistema operacional, armazenamento, aplicações em *deploy* e, possivelmente, algum controle limitado sobre os componentes de rede (como *host firewalls*).**"(MELL; GRANCE et al., 2011)

O modelo IaaS constitui a base para todos os demais modelos de serviço da nuvem. Através desse modelo, consumidores podem instanciar e administrar recursos virtualizados como máquinas, unidades de armazenamento de baixo nível (como armazenamento em blocos ou arquivos) e infraestrutura de rede (TANENBAUM; STEEN, 2007).

São alguns exemplos de IaaS:

- **Serviços de computação:** permitem o provisionamento de máquinas virtuais ou unidades computacionais virtualizadas sobre o hardware do provedor. Exemplos: Amazon Elastic Compute Cloud, Azure Virtual Machines e Google Compute Engine.
- **Serviços de armazenamento:** permite o provisionamento de unidades virtualizadas de armazenamento sobre o hardware do provedor. Alguns serviços dessa categoria tem como objetivo servir como sistema de backup e recuperação. Exemplos comerciais: Amazon EFS, Amazon EBS e Amazon Backup.
- **Serviços de rede:** permitem o provisionamento de recursos virtualizados de rede sobre o hardware do provedor. Fora as funcionalidades convencionais de rede (como roteamento, balanceadores de carga e firewalls), o provedor também pode oferecer redes de distribuição de conteúdo, entre outras capacidades. Exemplos comerciais: Amazon Citrix SD-WAN, Amazon Citrix ADC e Amazon CloudFront.

3.3.2 Plataforma como um Serviço (PaaS)

"**Serviço que provê ao consumidor a capacidade de realizar desenvolvimento e *deployment* de aplicações sobre uma infraestrutura da nuvem. As linguagens de programação, bibliotecas, serviços e ferramentas da aplicação devem**

ser compatíveis com o ambiente disponibilizado pelo provedor. Embora o consumidor não gerencie ou controle a infraestrutura subjacente da nuvem (como rede, servidores, sistemas operacionais ou armazenamento, ele tem controle sobre as aplicações em *deployment* e, possivelmente, controles de configuração para o ambiente hospedeiro da aplicação."(MELL; GRANCE et al., 2011)

Construído sobre IaaS, o modelo de serviço PaaS provê recursos para realizar o *deploy* de aplicações sem a necessidade de gerenciar a camada de Infraestrutura. Esse modelo de serviço também contempla ferramentas para desenvolver, testar, gerenciar o *deployment* e manipular elasticamente a escala de aplicações na nuvem (GOYAL; DADIZADEH, 2009).

Adicionalmente, temos associados a esse modelo, os serviços de armazenamento de objetos[2] e de bases de dados auto-gerenciadas (TANENBAUM; STEEN, 2007).

A incompatibilidade entre os *engines* de PaaS dos diferentes provedores de serviços na nuvem torna inviável a portabilidade dos *deployments* entre diferentes provedores. Essa é uma grande desvantagem na utilização desses *engines* para construir aplicações (GOYAL; DADIZADEH, 2009).

São alguns exemplos de PaaS:

- ***Deploy* de aplicações:** provê um ambiente virtualizado para execução e deployment de aplicações que possuam compatibilidade com o ambiente provido. Exemplos comerciais: Amazon Elastic Beanstalk e AWS Lambda [3];

- **Serviços de base de dados:** permite o provisionamento de bases de dados relacionais e não relacionais. Em geral, o consumidor dispõe de diversas facilidades de operação como gerência automatizada dos esquemas de replicação, clusterização de instâncias e gerenciamento de *backups*. Exemplos comerciais: Amazon RDS (SQL) e Amazon DynamoDB (NoSQL);

- **Plataformas de desenvolvimento e testes:** permitem a criação de plataformas virtualizadas para os diversos ambientes de execução (como desenvolvimento, testes, *staging* (ambiente de aceitação do usuário) e produção). Uma grande vantagem dessas plataformas é a facilidade obtida para definir e replicar ambientes de execução para as aplicações desenvolvidas. Exemplos comerciais: Amazon Cloud Development Kit e Google App Engine.

3.3.3 *Software* como um Serviço (SaaS)

"**Serviço que provê ao consumidor a capacidade de utilizar uma aplicação de um provedor sobre a infraestrutura de nuvem. As aplicações podem ser acessíveis a partir de clientes distintos (como um navegador Web ou alguma interface**

[2] Alguns serviços de nuvem, como o Amazon S3, podem atender às características de IaaS e PaaS, provendo desse modo ambos modelos de serviço, de acordo com o caso de uso (STEEN, 2021)

[3] Considera-se o modelo de Função como um Serviço (FaaS) um caso particular do modelo PaaS

de programa). Embora o consumidor não gerencie ou controle a infraestrutura subjacente da nuvem (como rede, servidores, sistemas operacionais, armazenamento e capacidades individuais da aplicação), ele tem controle sobre um conjunto restrito de configurações especificas de usuário da aplicação."(MELL; GRANCE et al., 2011)

Nesse modelo de serviço, o consumidor recebe um *software* pronto para ser utilizado pelo usuário final, sem necessidade da tratativa de questões de Infraestrutura e Plataforma. Na verdade, na maior parte dos casos, o consumidor nem mesmo possui acesso às configurações, a não ser aquelas estritamente relacionadas ao uso da aplicação, como preferências de usuários.

Para esse modelo de serviço, destaca-se duas vantagens importantes (BLOKDIJK; MENKEN, 2009):

- O fato de que as máquinas dos clientes funcionam como simples terminais de interface e todo processamento efetivo da aplicação ocorre na nuvem, implica na possibilidade de utilizar aplicações com cargas pesadas de trabalho, mesmo em máquinas com pouco poder computaional;

- Como consequência do fato anterior, o desacoplamento entre a carga de processamento e o terminal de interface com o usuário, permite que a aplicação seja executada pelo usuário em ambientes convencionais. A maioria das aplicações são acessíveis através de protocolos suportados pelos navegadores, sendo esse o seu ambiente de execução. Isso exime o usuário da necessidade de atualizar cada aplicação executada, sendo necessário apenas, manter uma versão de navegador Web minimamente atualizada.

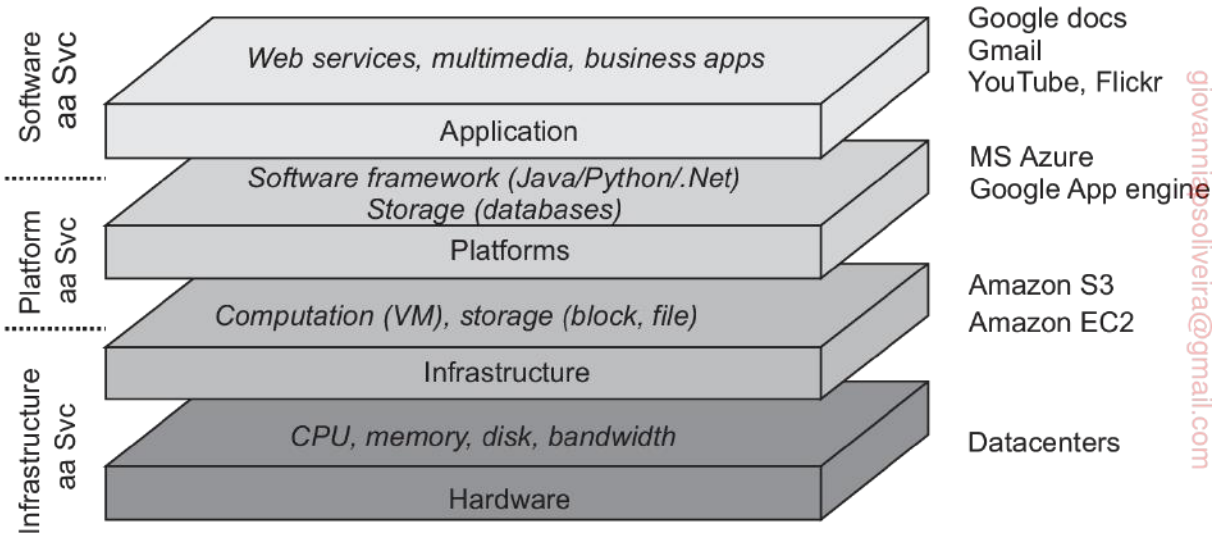
Aplicações servidas sob o modelo SaaS frequentemente são executadas em *multitenancy* (arquiteturas *multi-tenant*), isto é, a mesma instância de aplicação no provedor é compartilhada por diferentes *tenants* (grupos de clientes). A prática *multitenancy* é fortemente relacionada com o modelo de computação na nuvem (WIKIPEDIA, 2021a).

A organização e relação entre as camadas e os modelos de serviço tradicionais da nuvem, estão representados através da Figura 3.

Nesse modelo de serviço encontra-se a maior variedade de aplicações e ofertas de serviços. A seguir, exemplifica-se o modelo com algumas das aplicações de maior sucesso (MAYFIELD, 2016)(VLADIMIRSKIY, 2016)(NYANSA, 2016):

- **Salesforce**: Sendo a principal plataforma da atualidade para gerenciamento de relacionamento com o cliente, a Salesforce existe desde 1999 e foi uma das pioneiras no oferecimento de serviço sob o modelo SaaS (SALESFORCE, 2021b) (SALESFORCE, 2021a);

Figura 3 – Representação gráfica do modelo de camadas da computação na nuvem. Para cada modelo de serviço, à esquerda, há um conjunto de exemplos, à direita. O termo frontal dos paralelepípedos denomina a camada e o texto superior evidencia sua composição.



Fonte: (TANENBAUM; STEEN, 2007)

.

- **Google Workspace**: Antigamente conhecido como G Suite, a plataforma Google Workspace conta com diversas ferramentas SaaS de colaboração como: Gmail (cliente de e-mail), Calendar (gerência de agenda), Meet (chamadas de voz e vídeo), Chat (mensagens instantâneas), Drive (armazenamento), Docs (processador de texto), Sheets (planílhas eletrônicas), Slides (folhas de apresentação), Forms (formulários e enquetes) e Sites (gerenciamento de sítios na *web*) (GOOGLE, 2021);

- **Microsoft Office 365**: Oferta SaaS das aplicações do pacote Office (aplicações de escritório) e outras aplicações de colaboração como OneDrive (armazenamento), Skype e Teams (comunicação) (MICROSOFT, 2021);

- **Youtube e Netflix**: Plataformas para transmissão de mídia, como filmes, séries e documentários, sob-demanda. No caso do Youtube, é permitido aos usuários em geral fazer *upload* e distribuir conteúdo, tornando a plataforma também uma ferramenta de expressão social (YOUTUBE, 2021) (NETFLIX, 2021);

- **Shopify**: Plataforma para suporte à venda de produtos *online*. Possui ferramentas que vão desde simples componentes de venda integrados com provedores de pagamento até sistemas completos para operação de grandes negócios. O produto mais utilizado da plataforma é a ferramenta para criação, operação e gerência de lojas virtuais (SHOPIFY, 2021).

3.3.4 * como um Serviço (*aaS)

Uma década após a definição dos modelos de serviço básicos (IaaS, PaaS e SaaS), o acrônimo *aaS é utilizado para descrever especializações e combinações desses modelos. São alguns exemplos:

- **Mobile Backend como Serviço (MBaaS):** provê facilidades para integração de aplicações móveis com outros serviços da nuvem. Alguns exemplos de facilidades oferecidas por esse modelo de serviço são as integrações com serviços de notificação, armazenamento, autenticação, gerenciamento de usuários, redes sociais, relatório de *crashes* (falhas e bugs) e *analytics* (análise computacional sistemática de dados e estatísticas) (EVANGELIST, 2012)(WIKIPEDIA, 2021b);
- **Função como Serviço (FaaS):** provê uma plataforma onde os consumidores podem rodar funções sem provisionamento ou gerenciamento de servidores. Basicamente o consumidor apenas faz *upload* do código e o provedor toma todas as medidas necessárias para rodar o código com escalabilidade e alta-disponibilidade. O uso de FaaS é um modo de alcançar a arquitetura *serverless*, onde o provedor abstrai o servidor de execução do ambiente de desenvolvimento da aplicação (FOWLER, 2018).

3.4 MODELOS DE DEPLOYMENT

O modelo de computação na nuvem pode ser implantado em diversas configurações, caracterizadas pelos modelos de *deployment* apresentados nessa seção.

3.4.1 Nuvem pública

"**A infraestrutura de nuvem é provisionada para ser utilizada pelo público geral. Infraestruturas desse tipo podem ser possuídas, gerenciadas e operadas por organizações comerciais, acadêmicas, governamentais ou uma combinação entre organizações de diferentes natureza. Nesse modelo, o *deployment* é sob premissas (*on-premise*[4]) do provedor da nuvem.**"(MELL; GRANCE et al., 2011)

São nuvens cuja infraestrutura de TI é *off-premise*[5] do consumidor. A comunicação entre o provedor e o consumidor ocorre geralmente através da Internet (Figura 4).

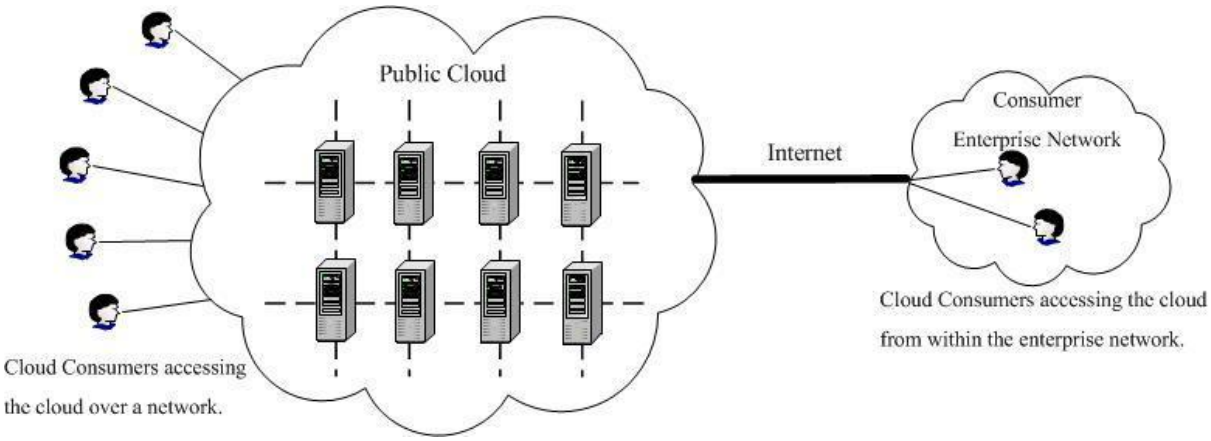
Adicionalmente, os provedores públicos também oferecem serviços que permitem realizar *deployment* de suas respectivas nuvens, sobre infraestrutura on-premise do consumidor (AWS, 2021a).

Os principais provedores públicos de computação na nuvem da atualidade são (STATISTA, 2021b)(RENO, 2021):

[4] On-premise é a condição de uma instituição ou organização que possui e gerencia sua própria infraestrutura de hardware

[5] Off-premise tem sentido oposto a on-premise. É a condição de uma instituição ou organização que utiliza a infraestrutura de hardware terceiros.

Figura 4 – Esquema básico de comunicação entre o provedor e o consumidor em nuvens públicas.



Fonte: (LIU et al., 2011)

.

- Amazon Web Services (AWS);
- Google Cloud Platform (GCP);
- IBM Cloud;
- Alibaba Cloud;
- Microsoft Azure.

3.4.2 Nuvem privada

"**A infraestrutura de nuvem é provisionada para uso exclusivo de uma única organização com múltiplos consumidores. Infraestruturas desse tipo podem ser possuídas, gerenciadas e operadas pela própria organização, terceiros, ou uma combinação entre as duas alternativas. Nesse modelo, o *deployment* pode se dar *on-premise* ou *off-premise* do consumidor.**"(MELL; GRANCE et al., 2011)

Uma nuvem privada provê infraestrutura e recursos computacionais exclusivos para consumidores da mesma organização. A comunicação entre provedor e consumidores se dá, geralmente, através de redes de longa distância (WANs) (RUPARELIA, 2016).

Pode-se identificar dois subtipos de nuvens privadas (REDHAT, 2021a):

- **Nuvens privadas gerenciadas:** Os consumidores criam e usam uma nuvem privada implantada, configurada e gerenciada por uma terceira parte. Isso permite que empresas possam usufruir dos benefícios de uma nuvem privada, sem a necessidade de possuir os recursos humanos necessários para gerenciá-la.

- **Nuvens dedicadas:** Uma nuvem rodando sobre a infraestrutura de outra nuvem privada ou pública.

As Figuras 5 e 6 evidenciam a diferença entre os modelos de implantação *on-premise* e *off-premise* do consumidor para o modelo privado de deployment. No caso do deployment *on-premise*, a infraestrutura de nuvem encontram-se no domínio do consumidor.

3.4.3 Nuvem comunitária

"**A infraestrutura de nuvem é provisionada para uso exclusivo de uma comunidade de consumidores de organizações com interesses comuns. Infraestruturas desse tipo podem ser possuídas, gerenciadas e operadas por uma ou mais organizações da comunidade, terceiros, ou uma combinação entre as duas alternativas. Nesse modelo, o *deployment* pode se dar *on-premise* (Figura 7) ou *off-premise* (Figura 8) do consumidor.**"(MELL; GRANCE et al., 2011)

É uma extensão do conceito de nuvens privadas, que agrupa o corpo de consumidores de organizações distintas com interesses em comum (ex.: QTS Healthcare Community Cloud (QTS, 2016)).

Outra diferença em relação à nuvens privadas é que a comunicação entre o provedor e os consumidores se dá, geralmente, através da Internet (RUPARELIA, 2016).

Nuvens comunitárias podem surgir da combinação de nuvens privadas de cada uma das organizações (ver Figura 7) ou da utilização de uma infraestrutura totalmente *off-premise* (Figura 8).

3.4.4 Nuvem híbrida

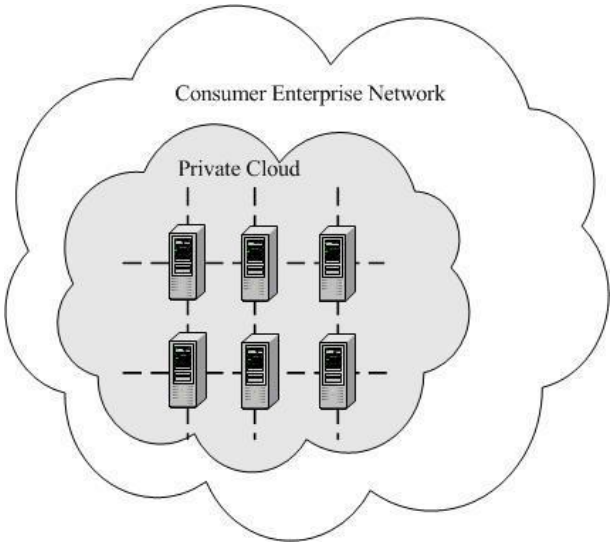
"**A infraestrutura de nuvem é uma composição de dois ou mais tipos distintos de infraestrutura de nuvem (privada, comunitária ou pública) que permanecem como entidades únicas (ver Figura 9), mas estão associadas através de tecnologias (padronizadas ou proprietárias) que proporcionam portabilidade de dados e aplicações.**"(MELL; GRANCE et al., 2011)

O termo "multi nuvens" refere-se a nuvens compostas pelo agrupamento de duas ou mais nuvens dentro do mesmo modelo de deployment (apenas público, apenas privado ou apenas comunitário).

O termo "nuvem híbrida" refere-se a nuvens compostas pelo agrupamento de mais de uma nuvem em modelos de deployment distintos (público, privado e comunitário). A integração dos serviços oferecidos pelos diferentes provedores ocorre através do uso de mecanismos de orquestração de cargas de trabalho (REDHAT, 2021a).

A interconexão entre os diversos ambientes que compõem a nuvem híbrida podem ocorrer através de redes locais (LANs), redes de longa distância (WANs), redes virtuais privadas (VPNs) e APIs (REDHAT, 2021c).

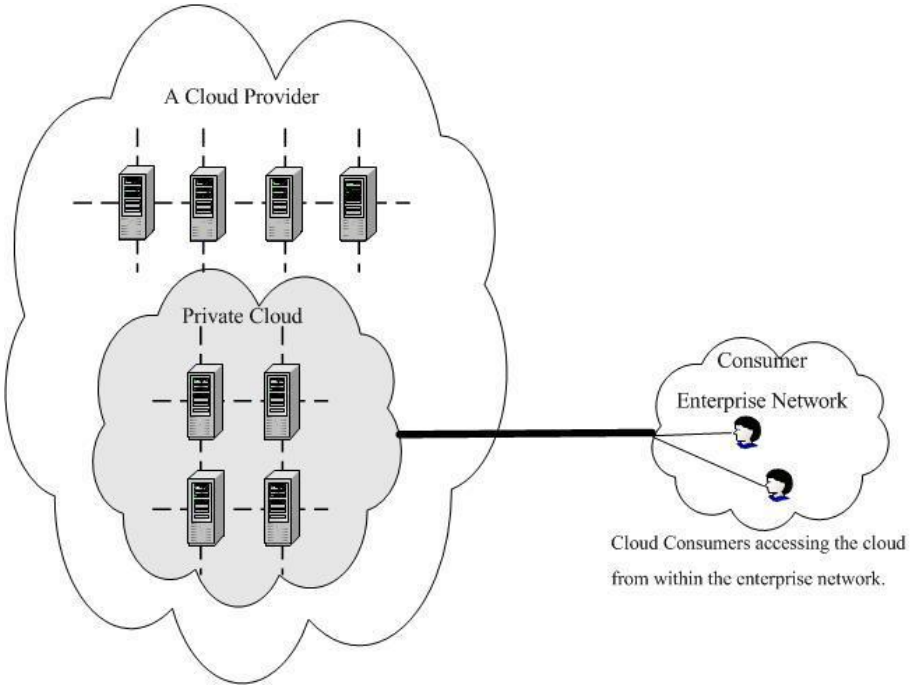
Figura 5 – Esquema básico de comunicação entre o provedor e o consumidor em uma nuvem privada implantada sobre infraestrutura on-premise do consumidor.



Fonte: (LIU et al., 2011)

.

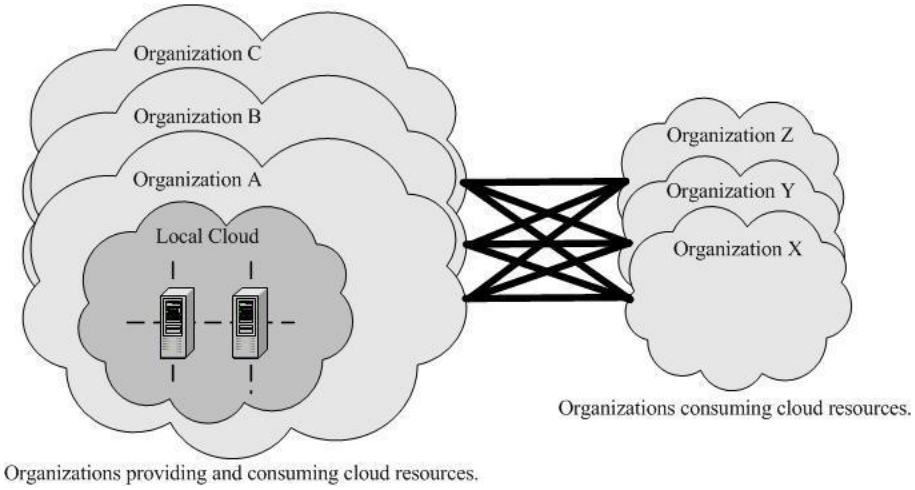
Figura 6 – Estrutura básica de comunicação entre o provedor e o consumidor em uma nuvem privada implantada sobre infraestrutura off-premise do consumidor.



Fonte: (LIU et al., 2011)

.

Figura 7 – Esquema básico de comunicação entre os provedores e os consumidores em nuvens comunitárias implantadas em infraestrutura on-premise do conjunto de organizações mantenedoras.

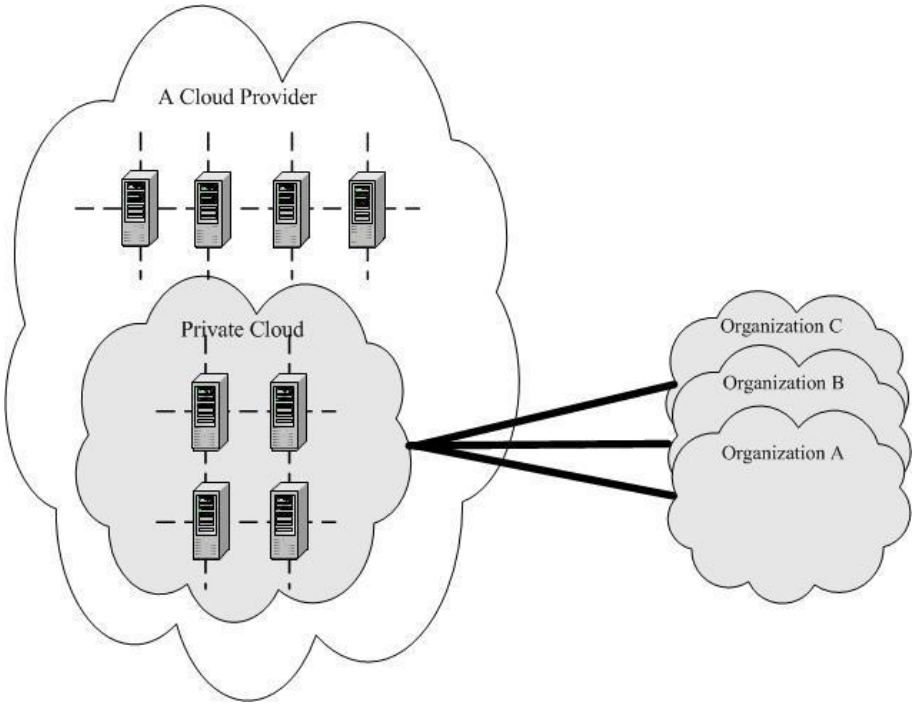


Fonte: (LIU et al., 2011)

.

A nuvem híbrida oferece aos consumidores uma interface unificada com orquestração para consumo de serviços de diferentes provedores (Figura 9).

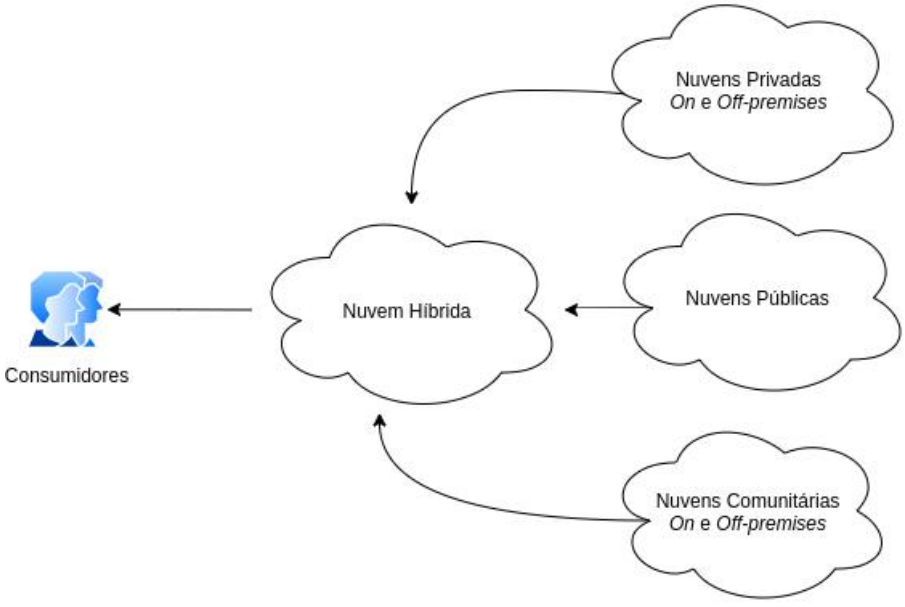
Figura 8 – Esquema básico de comunicação entre o provedor e os consumidores em nuvens comunitárias implantadas na infraestrutura do provedor.



Fonte: (LIU et al., 2011)

.

Figura 9 – Esquema básico de interface entre provedor e consumidor em nuvens híbridas.



4 SÍNTESE HISTÓRICA DA COMPUTAÇÃO NA NUVEM

Nesse capítulo, apresenta-se uma visão concisa e sintética dos principais acontecimentos históricos que estiveram diretamente relacionados ao desenvolvimento e consolidação do modelo de computação na nuvem.

4.1 IDEALIZAÇÃO

Uma das primeiras proposições do modelo de computação na nuvem ocorreu através do discurso de John McCarthy, em 1961, no evento comemorativo do centenário do Instituto de Tecnologia de Massachusetts (MIT) (GARFINKEL, 1999):

" A computação pode em algum dia ser organizada como uma utilidade pública, assim como o telefone é uma utilizade pública ... A computação utilitária poderia se tornar base de uma nova e importante indústria. "

Em seu artigo de nome "Os Computadores de Amanhã", publicado em 1964, Martin Greenberger realiza uma comparação entre o fornecimento de energia elétrica e sua ideia de computação utilitária. Alguns dos pontos mais importantes de sua analogia são (GREENBERGER, 1964):

- Alta disponibilidade é uma característica desejada para o modelo de computação utilitária, que se tornava cada vez mais popular, a medida em que os computadores se tornavam mais confiáveis. As sobrecargas de demanda pontuais no tempo já era um problema visado;
- A implementação se tornaria cada vez mais possível a medida que o custo associado à computação fosse reduzido;
- A uniformidade do recurso provido e distribuído pela indústria energética tornava mais simples sua distribuição (principalmente para múltiplos consumidores) em comparação com os recursos a serem oferecidos pela computação utilitária;
- O custo para gerenciar o acesso de múltiplos consumidores à mesma infraestrutura ainda era muito alto para o conjunto de tecnologias disponíveis na época.

Um dos primeiros exemplos da industrialização da computação utilitária se deu a partir da tecnologia proprietária IBM Remote Job Entry (RJE). Essa tecnologia permitia que múltiplos usuários submetessem tarefas, a partir de terminais, para serem processadas por um *mainframe* com escalonamento de tempo compartilhado. (IBM, 1968).

4.2 IMPLEMENTAÇÃO

Embora idealizada desde 1961, a computação utilitária, na forma de computação na nuvem, demorou 30 anos para começar a ser implementada, em decorrência das carências tecnológicas existentes (ver Capítulo 2).

Durante os anos 90, companias de telecomunicações, que antes ofereciam circuitos dedicados para comunicação ponto-a-ponto, passaram a oferecer Redes Privadas Virtuais (VPNs) como serviço com Qualidade de Serviço (QoS) comparável a dos circuitos dedicados e com custo inferior. O símbolo de nuvem passou a ser usado para representar a infraestrutura que conectava o consumidor ao provedor de serviços.

Referências ao termo computação na nuvem com a semântica atual tiveram sua aparição em documentos internos da Compaq em 1996.

Após a consolidação da virtualização no final dos anos 90, em associação com tecnologias de computação em grade já estabelecidas, tornou-se possível a implementação de nuvens privadas e híbridas (GRIFFIN, 2018).

A difusão da Internet, principalmente a partir da popularização da banda larga nos anos 2000 (MURRAY-WEST, 2016), viabilizou os *deployments* baseados em nuvens públicas.

4.3 POPULARIZAÇÃO

O modelo de computação na nuvem tornou-se extremamente popular após a consolidação dos modelos de *deployment* em nuvens públicas, usando a Internet como principal rede de comunicação entre os provedores e o consumidores de serviços.

Um dos grandes marcos comerciais na história da computação na nuvem se deu com o lançamento dos serviços Amazon Simple Storage Service (Amazon S3) e Amazon Elastic Compute Cloud (Amazon EC2), no ano de 2006, tornando públicas algumas das soluções implementadas para atender as necessidades relativas à infraestrutura da distribuição da aplicação Web de vendas (AWS, 2006b).

Ainda em 2006, o Google lançou sua linha de serviços Google Docs com o Google Spreadsheet e Writely que proviam programas de escritório como aplicações *multitenant* rodando sobre uma infraestrutura de nuvem (MARSHALL, 2006).

No início de 2010, a Microsoft lançou a plataforma de computação na nuvem Windows Azure (renomeada futuramene para Microsoft Azure). Os primeiros serviços oferecidos pela plataforma foram o Windows Azure e SQL Azure, provendo máquinas virtuais e bases de dados rodando sobre a nuvem, respectivamente (SRIVASTAVA, 2010) (YI, 2011).

Ainda em 2010, a NASA em conjunto com a Rackspace Hosting lançaram uma plataforma de computação na nuvem de código aberto chamada OpenStack (CURRY, 2010). Atualmente, o OpenStack é uma plataforma utilizada para construir e gerenciar nuvens privadas (e públicas) a partir de *pools* de recursos virtualizados. Por meio dos *Projects*

(módulos operacionais) que compõem a plataforma, são providos serviços essenciais de computação, rede, armazenamento, identidade e imagem na nuvem (REDHAT, 2021b).

Embora a plataforma Google Cloud tenha sido lançada oficialmente em 2008 com a PaaS Google App Engine (MCDONALD, 2008), um serviço para computação só se tornou disponível na plataforma em 2013, com o lançamento do Google Compute Engine (DARROW, 2013).

Em 2011, o NIST publicou uma série de artigos a fim de definir de forma clara a computação na nuvem, bem como sua arquitetura (MELL; GRANCE et al., 2011) (LIU et al., 2011).

5 COMPARAÇÃO ENTRE PROVEDORES PÚBLICOS DE COMPUTAÇÃO NA NUVEM

Nesse capítulo é realizada a análise de três provedores públicos de computação na nuvem, escolhidos de acordo com o critério de popularidade. Para isso, são considerados dados de faturamento e estatísticas de uso.

Após a seleção, os provedores são avaliados a partir dos seguintes parâmetros:

- Variedade da oferta de serviços;
- Disposição geográfica da infraestrutura de nuvem;
- Disponibilidade, baseada em histórico de *outages* e Acordos de Nível de Serviço (SLAs);
- Desempenho real da infraestrutura de rede dos provedores;
- Modelos de tarifação e valor de custo absoluto (considerando um caso de uso como exemplo);
- Conformidade com padrões e certificações de qualidade.

Em decorrência da grande variedade de serviços oferecidos por cada um dos provedores, apenas um sub-conjunto de maior relevância é analisado.

A ideia é de que, no final da análise, seja possível classificar competitivamente os provedores em cada um dos aspectos mencionados, ainda que não seja possível estabelecer uma ordem global para todos os aspectos[1].

5.1 MERCADO E ACEITAÇÃO

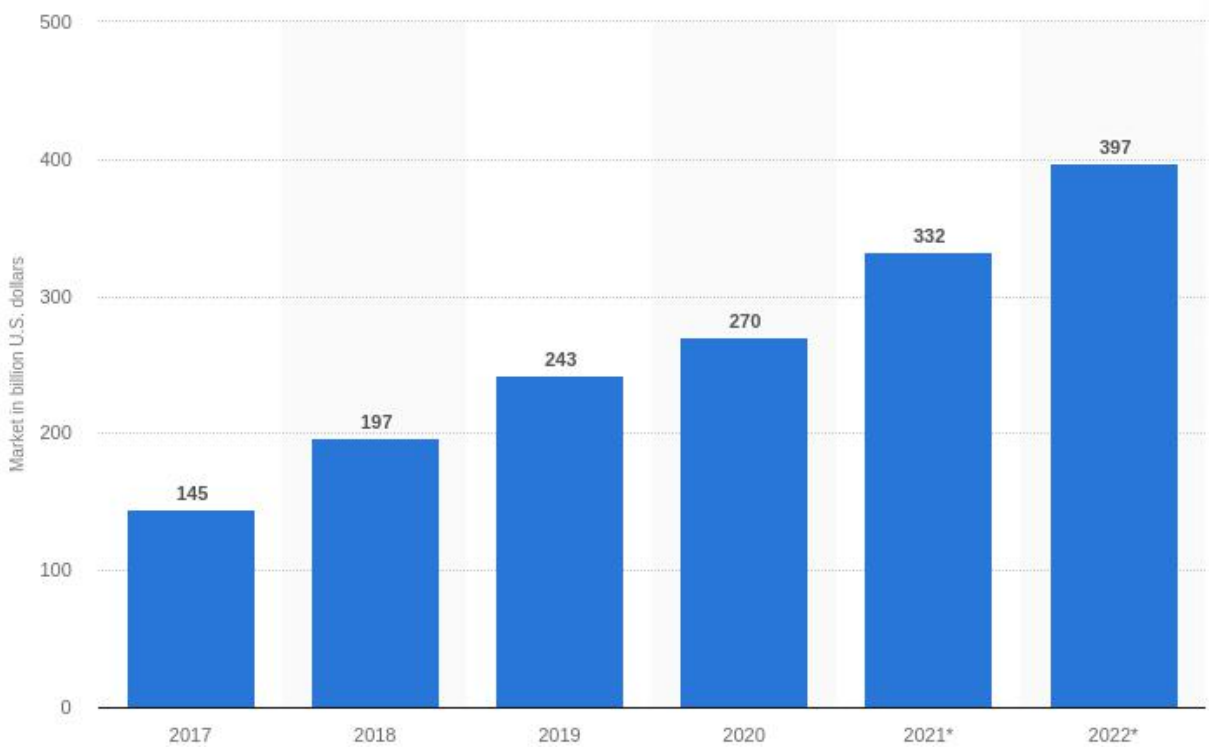
Analisando os gastos do usuário final, percebe-se que o mercado gerado pela oferta de serviços em nuvens públicas tem crescido monotonamente, com alta taxa de crescimento (Figura 10). Desse modo, o modelo de computação na nuvem tem ganhado cada vez mais notoriedade no mercado de serviços de computação e na economia global, em geral.

Considerando a estimativa de paticipação no mercado de cada um dos provedores públicos, nas modalidades de Iaas e PaaS, no quarto trimestre de 2020 (Figura 11), obtem-se os provedores de maior procentagem: AWS ($\approx$ 32 %), Azure ($\approx$ 20 %) e GCP ($\approx$ 9 %).

Considerando a estimativa de adoção dos provedores por parte das instituições privadas, nota-se que, embora a ordem encontrada na estatística de participação de mercado

[1] A análise de cada aspecto pode conter metodologias particulares, definidas de forma explícita ou implicita nas seções correspondentes às análises

Figura 10 – Estimativa de gastos de usuários finais com serviços em nuvens públicas, em bilhões de dolares, de 2017 a 2022. * o valor exibido para os anos de 2021 e 2022 são inferidos a partir dos dados coletados de 2017 a 2020



Fonte: (STATISTA, 2021b)

.

se mantenha, há uma aproximação entre os percentuais obtidos para a AWS (≈ 79%) e Azure (≈ 76%).

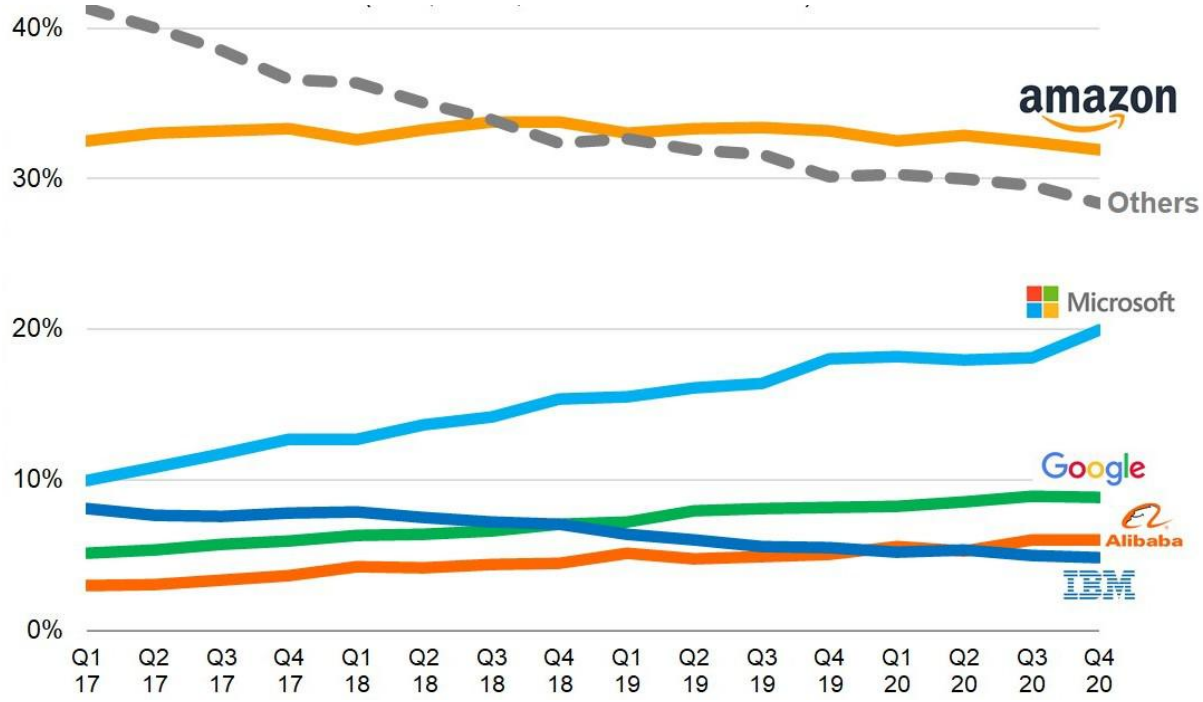
O alto percentual de empresas experimentando a GCP (≈ 23%) indica um possível crescimento no número de usuários em curto prazo, mas a plataforma ainda se mantém atrás das duas primeiras concorrentes, com ≈ 49% (aproximandamente 30% a menos que AWS e 27% a menos que Azure).

Desse modo, seleciona-se pelo critério de relevância no mercado os provedores públicos: AWS, Azure e GCP.

A análise dos dados da Figura 12 também permite identificar uma tendência contemporânea dos consumidores de serviços na nuvem: o uso de multi-clouds.

A Figura 13 apresenta os dados obtidos para categorização do uso de serviços na nuvem, de acordo com o tipo de *deployment*. O resultado da análise é extremamente polarizado: 92% das empresas que fazem uso de serviços de computação na nuvem, utilizam multi-nuvem. Em especial, 82% das empresas entrevistadas pela pesquisa, fazem uso de nuvens híbridas.

Figura 11 – Estimativas trimestral da porcentagem de participação de mercado dos principais provedores de computação na nuvem, de 2017 a 2020.



Fonte: (RENO, 2021)

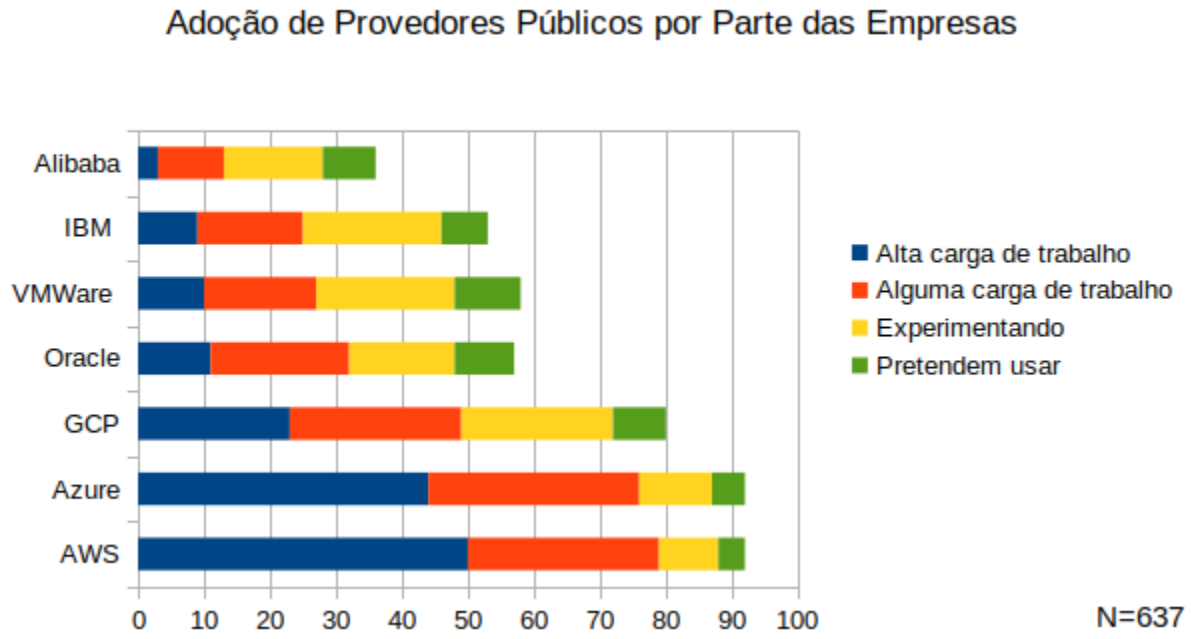
.

5.2 VARIEDADE E OFERTA DE SERVIÇOS

Os conjunto de serviços oferecidos pelos provedores analisados possuem similaridades, e, em diversos casos, é possível estabelecer uma relação de equivalência entre os serviços de diferentes provedores. A Tabela 1 indexa os seguintes serviços: Servidor Virtual, Generenciamento de Contêineres, Ambientes para Desenvolvimento e *Deployment* de Aplicações, FaaS, Armazenamento de Objetos, Armazenamento de Blocos, Armazenamento de Arquivos, Gerenciamento de Bases de Dados Relacionais, Gerenciamento de Bases de Dados Não-Relacionais, Redes de Entrega de Conteúdo (CDNs) e Gerenciamento de Rotas e Tráfego.

Com relação ao uso dos serviços segregados por modelo, de acordo com a estatística apresentada na Figura 14, compreende-se que, a maior parte dos gastos do usuário final com a nuvem se dá com SaaS, seguido de IaaS e PaaS, decrescentemente, nessa ordem. Para os modelos de serviço analisados, tem-se as porcentagens de utilização em 2021: SaaS (46.62 %), PaaS (22.55 %) e IaaS (30.83 %).

Figura 12 – Estimativas de adoção de nuvens públicas por empresas em 2021 (obtido com espaço amostral de 637 companhias de características heterogêneas).



Fonte: Adaptado de (FLEXERA, 2021)

.

Figura 13 – Distribuição dos tipos de deployment de computação na nuvem utilizados pelas empresas participantes da pesquisa.

Fonte: Adaptado de (FLEXERA, 2021)

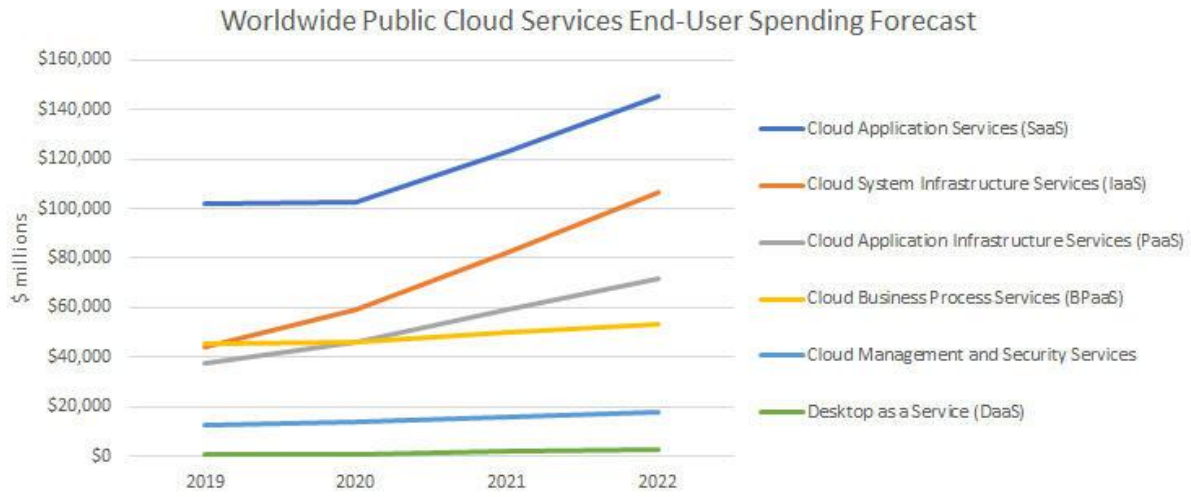
.

Quadro 1 – Equivalência de alguns dos principais serviços oferecidos pelos provedores públicos

Serviço	**AWS**	**Azure**	**GCP**
Servidor Virtual	Amazon Elastic Compute Cloud (EC2)	Azure Virtual Machine	Google Compute Engine
Gerenciamento de Contêineres	Amazon EC2 Container Service e Amazon Elastic Kubernetes Service (EKS)	Azure Container Instances e Azure Kubernetes Services (AKS)	Google Kubernetes Engine
Ambientes para Desenvolvimento e *Deployment* de Aplicações	AWS Elastic Beanstalk	Azure Web Apps	Google App Engine
FaaS	AWS Lambda	Azure Functions	Google Cloud Functions
Armazenamento de Objetos	Amazon Simple Storage Service (S3)	Azure Blob	Google Cloud Storage
Armazenamento de Blocos	Amazon Elastic Block Storage (EBS)	Azure Page Blobs e Azure Managed Disks	Google Persistent Disk
Armazenamento de Arquivos	Amazon Elastic File Storage (EFS)	Azure File Storage	Google File Store
Gerenciamento Bases de Dados Relacionais	Amazon Aurora e RDS	Azure SQL Database e versões Postgre e MySQL	Google Cloud SQL e Spanner
Gerenciamento Bases de Dados Não-Relacionais	Amazon DynamoDB, DocumentDB e Neptune	Azure CosmosDB, Table Storage e Time Series Insights	Google Cloud Datastore e BigTable
Rede de Entrega de Conteúdo (CDN)	Amazon CloudFront	Azure CDN	Google Cloud CDN
Gerenciamento Rotas e Tráfego	Amazon Route 53	Azure Traffic Manager	Cloud Load Balancer + Cloud DNS

Fonte: (COMPARECLOUD.IN, 2021)

Figura 14 – Estimativa dos gastos do usuário final com serviços oferecidos em nuvens públicas, segregados por tipo de serviço, em milhões de dólares.



Fonte: (MCLELLAN, 2021)

.

5.3 DISPOSIÇÃO GEOGRÁFICA DA INFRAESTRUTURA DE NUVEM

A proximidade geográfica entre provedor e consumidor de serviços na nuvem pode surtir efeitos signifcativos sobre a latência da comunicação, devido a fatores naturais dos meios de comunicação. A medida que que a distância cresce, também cresce, proporcionalmente, o tempo de propagação dos sinais que trafegam sobre esses meios (ex.: fio metálico, atmosfera e fibra ótica) (KUROSE, 2005) [2].

Com o propósito de tornar-se viável para consumidores de todo o mundo e oferecer um serviço global, os provedores da nuvem dividem sua infraestrutura em regiões. A definição de uma região pode variar de acordo com o provedor. A seguir, uma adaptação da definição fornecida pela AWS (AWS, 2021s):

Definição 4 ***: Regiões****, no contexto de infraestrutura de computação na nuvem, são localidades físicas ao redor do mundo onde os provedores clusterizam seus data centers. Dentro de cada Região, os data centers são agrupados logicamente, formando uma ou mais Zonas de Disponibilidade. O conjunto de múltiplas Zonas de Disponibilidade isoladas e fisicamente separadas compõe a Região.*

Define-se o conceito de Zona de Disponibilidade a partir da mesma fonte literária (AWS, 2021s):

Definição 5 ***: Zonas de Disponibilidade (AZ)****, no contexto de infraestrutura de computação na nuvem, são sub-regiões compostas por um ou mais data centers com energia elétrica, rede e conectividade redundantes. As AZs que compõem uma Região são interligadas através de canais redundantes, de alta velocidade e baixa latência. Tipicamente, AZs distam geograficamente umas das outras para aumentar a eficiência na prevenção às falhas de disponibilidade.*

AZs são utilizadas frequentemente para implementar esquemas de redundância e balanceamento de carga, simplificando o processo para desenvolver e implantar um componente, serviço ou aplicação com resiliência e alta disponibilidade.

Contudo, é necessário reiterar que essas definições podem variar de acordo com o provedor, o que torna imprecisa a realização de uma comparação direta entre os provedores, através do número absoluto de regiões ou AZs.

Um outro fator que inviabiliza a comparação quantitativa de AZs é a variação de disponibilidade de serviços nas AZs. Nem todas AZs contêm todos os serviços do provedor, implicando na variação de entrega de valor para o consumidor.

[2] O paradigma *Edge Computing* baseia-se nesse princípio e propõe uma arquitetura onde os elemento de comunicação estão mais próximos dos clientes ou das fontes de dados, com o intuito de reduzir a latência e o volume de comunicação entre os componentes do sistema distribuído (WIKIPEDIA, 2021d)

Nas subseções a seguir, é realizada uma contagem de acordo com as definições apresentadas anteriormente, mas sem considerar a variação de disponibilidade de serviços nas AZs. Adicionalmente, realiza-se um teste prático para obter a lista de regiões disponíveis para o serviço de computação padrão de cada um dos provedores (Amazon EC2, Azure VM e Google Cloud Compute Engine).

5.3.1 AWS

Resumo da infraestrutura global provida pela AWS:

- **Regiões disponíveis:** 25
- **Zonas de disponibilidade:** 80
- **Países e territórios atendidos:** 245

Fonte: (AWS, 2021r)

Usando a AWS CLI, obteve-se a lista de Regiões para as quais está disponível o serviço EC2, totalizando 21 Regiões, são elas: ['af-south-1', 'eu-north-1', 'ap-south-1', 'eu-west-3', 'eu-west-2', 'eu-south-1', 'eu-west-1', 'ap-northeast-3', 'ap-northeast-2', 'me-south-1', 'ap-northeast-1', 'sa-east-1', 'ca-central-1', 'ap-east-1', 'ap-southeast-1', 'ap-southeast-2', 'eu-central-1', 'us-east-1', 'us-east-2', 'us-west-1', 'us-west-2'].

Figura 15 – Mapa de regiões de disponibilidade atuais e previstas pela AWS.

Fonte: (AWS, 2021r)

.

5.3.2 Azure

Resumo da infraestrutura global provida pela Azure:

- **Regiões disponíveis:** 46
- **Zonas de disponibilidade:** Não identificado
- **Países e territórios atendidos:** 140+

Fonte: (AZURE, 2021a)

Usando a Azure CLI, obteve-se a lista de Regiões para as quais está disponível o serviço VM, totalizando em 42 Regiões, são elas: ['South Africa North', 'Germany West Central', 'Australia Central', 'Australia East', 'Australia Southeast', 'US DoD Central', 'US DoD East', 'US Gov Arizona', 'US Gov Texas', 'US Gov Virginia', 'Brazil South', 'Canada Central', 'Canada East', 'China East', 'China East 2', 'China North', 'China North 2', 'Korea Central', 'Korea South', 'UAE North', 'Central US', 'East US', 'East US 2', 'North Central US', 'South Central US', 'West Central US', 'West US', 'West US 2', 'North Europe', 'West Europe', 'France Central', 'Central India', 'South India', 'West India', 'Japan East', 'Japan West', 'Norway East', 'East Asia', 'Southeast Asia', 'UK South', 'UK West', 'Switzerland North']

Figura 16 – Mapa de regiões e zonas de disponibilidade atuais e previstas pela Azure.

Fonte: (AZURE, 2021a)

.

5.3.3 GCP

Resumo da infraestrutura global provida pela GCP:

- **Regiões disponíveis:** 24
- **Zonas de disponibilidade:** 73
- **Países e territórios atendidos:** 200+

Fonte: (GCP, 2021p)

Usando a GCP CLI, obteve-se a lista de Regiões para as quais está disponível o serviço Compute Engine, totalizando em 25 Regiões, são elas: ['us-east1', 'us-east4', 'us-central1', 'us-west1', 'europe-west4', 'europe-west1', 'europe-west3', 'europe-west2', 'asia-east1', 'asia-southeast1', 'asia-northeast1', 'asia-south1', 'australia-southeast1', 'southamerica-east1', 'asia-east2', 'asia-northeast2', 'asia-northeast3', 'asia-southeast2', 'europe-central2', 'europe-north1', 'europe-west6', 'northamerica-northeast1', 'us-west2', 'us-west3', 'us-west4']

Figura 17 – Mapa de regiões de disponibilidade, atuais e previstas pela GCP.

Fonte: (GCP, 2021p)

.

5.4 DISPONIBILIDADE

Nessa seção, analisa-se a disponibilidade dos provedores sob a ótica dos Acordos de Níveis de Serviço (SLAs) e registros de *outages* (contando com casos de indisponibilidade total ou estado de degradação parcial do serviço).

5.4.1 SLAs

Os dados apresentados nessa subseção foram obtidos dos documentos oficiais, disponibilizados pelos provedores (AWS, 2021p)(AZURE, 2021c)(GCP, 2021j).

As garantias dos serviços de computação oferecidos pelos provedores públicos são tipicamente formalizadas por meio de Acordos de Nível de Serviço (SLAs). Os provedores podem oferecer SLAs distintos, inclusive para o mesmo serviço, conforme será demonstrado nesta seção.

No caso do não cumprimento do fornecimento do serviço segundo valores pré-estabelecidos pelo SLA, o consumidor recebe parte do seu dinheiro de volta na forma de crédito para futuras tarifações aplicadas pelo provedor. A porcentagem de crédito estornado é inversamente proporcional à porcentagem de disponibilidade do serviço (*Disponibilidade*) durante o ciclo mensal de cobrança do respectivo consumidor. O montante estornado é calculado aplicando a porcentagem de crédito obtida, sobre o valor total de tarifação do respectivo serviço, na respectiva região onde o SLA foi violado.

A seguir, são indexados os SLAs disponíveis para cada um dos provedores analisados, considerando os principais serviços disponíveis para as categorias de computação, armazenamento, rede de distribuição de conteúdo (CDN), bases de dados relacionais (SQL) e não-relacionais (NoSQL):

5.4.1.1 Computação

Para a avaliação dos SLAs disponíveis para instâncias virtualizadas de computação na nuvem, considera-se dois casos: disponibilidade do serviço com redundância, composto por instâncias clusterizadas (Tabela 2), e disponibilidade de cada instância isolada (Tabela 3).

Na análise de instâncias com redundância, os provedores oferecem as mesmas garantias de níveis de serviço.

Na análise de instâncias isoladas, os provedores igualam a porcentagem estornada para o caso limite (disponibilidade $\leq 90\%$). Entretanto, os provedores GCP e Azure possuem maior discretização antes do caso limite. Dentre os provedores analisados, o GCP oferece o melhor negócio da perspectiva de SLA.

Quadro 2 – SLAs oferecidos pelos provedores analisados para serviços de computação em instâncias redundantes.

Provedor	**Serviço**	***Disponibilidade*** (U_t)	**% estornada em crédito**
AWS	Amazon EC2	$99\% < U_t \leq 99.99\%$	10%
		$95\% < U_t \leq 99\%$	30%
		$U_t \leq 95\%$	100%
Azure	Azure VM	$99\% < U_t \leq 99.99\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%
GCP	Google Compute Engine	$99\% < U_t \leq 99.99\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%

Fonte: (AWS, 2021d)(AZURE, 2020b)(GCP, 2021d)

Quadro 3 – SLAs oferecidos pelos provedores analisados para serviços de computação em cada instância isolada. Excepcionalmente nesse caso, os provedores AWS e Azure utilizam o minuto como base de tempo para o cálculo do parâmetro de *Disponibilidade*.

Provedor	**Serviço**	***Disponibilidade*** (U_t)	**% estornada em crédito**
AWS	Amazon EC2	$90\% \leq U_t$	100%
Azure	Azure Virtual Machine ver. Std. HDD	$92\% < U_t \leq 95\%$	10%
		$90\% < U_t \leq 92\%$	25%
		$U_t \leq 90\%$	100%
GCP	Google Compute Engine	$95\% < U_t \leq 99.5\%$	10%
		$90\% < U_t \leq 95\%$	25%
		$U_t \leq 90\%$	100%

Fonte: (AWS, 2021d)(AZURE, 2020b)(GCP, 2021d)

5.4.1.2 Armazenamento

Na categoria de armazenamento, observa-se que os melhores níveis de garantia são oferecidos pela GCP, seguida da AWS e, por fim, Azure.

Observa-se que as faixas de desconto iniciam-se em 99.95% para a GCP e em 99.9% para a AWS e Azure. Embora esses valores possam parecer relativamente próximos, entende-se através do cálculo a seguir que as diferenças são extremamente relevantes:

Para o caso do nível de 99.9%, o consumidor poderá contar com até 74 horas e 24 minutos (0.1% do tempo do ciclo mensal) de indisponibilidade, sem ter o SLA violado. Para o caso do nível de 99.95%, o consumidor poderá contar com até 37 horas e 6 minutos de indisponibilidade. Indicando que a plataforma GCP oferece duas vezes mais disponibilidade em relação aos concorrentes analisados, para o nível de SLA em questão.

Quadro 4 – SLAs oferecidos pelos provedores analisados para armazenamento na nuvem.

Provedor	**Serviço**	***Disponibilidade*(U_t)**	**% estornada em crédito**
AWS	Amazon S3	$99\% < U_t \leq 99.9\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%
Azure	Azure Storage Acc. Hot Blob	$99\% < U_t \leq 99.9\%$	10%
		$U_t \leq 99\%$	25%
Azure	Azure Storage Acc. Cool Blob	$98\% < U_t \leq 99\%$	10%
		$U_t \leq 98\%$	25%
GCP	Google Cloud Storage	$99\% < U_t \leq 99.95\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%

Fonte: (AWS, 2021f)(AZURE, 2019)(GCP, 2021c)

5.4.1.3 Redes de Entrega de Conteúdo (CDNs)

Para os serviços de CDN, indentifica-se que o melhor acordo é oferecido pela GCP, seguida de AWS e Azure, nessa ordem. O argumento é, novamente, um SLA superior ao dos concorrentes no melhor caso. Diferente dos concorrentes analisados, a Azure não oferece desconto total no pior caso, tornando sua oferta de SLA inferior para esse serviço.

Quadro 5 – SLAs oferecidos pelos provedores analisados para CDNs na nuvem.

Provedor	**Serviço**	***Disponibilidade*(U_t)**	**% estornada em crédito**
AWS	Amazon CloudFront Service	$99\% < U_t \leq 99.9\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%
Azure	Azure CDN	$99.5\% < U_t \leq 99.9\%$	10%
		$U_t \leq 99.5\%$	25%
GCP	Google Cloud CDN	$99\% < U_t \leq 99.95\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%

Fonte: (AWS, 2021c)(AZURE, 2015)(GCP, 2021a)

5.4.1.4 Bases de dados relacionais

Na categoria de bases de dados, o provedor com melhor acordo é a Azure (oferecendo uma disponibilidade de 99.995% para a primeira faixa de desconto para a versão Business Critical), seguida da GCP (com 99.99%) e da AWS (com 99.9%).

Quadro 6 – SLAs oferecidos pelos provedores analisados para bases de dados relacionais na nuvem.

Provedor	**Serviço**	***Disponibilidade*(U_t)**	**% estornada em crédito**
AWS	Amazon Aurora	$99\% < U_t \leq 99.9\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%
Azure	Azure SQL DB Business Critical	$99\% < U_t \leq 99.995\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%
Azure	Azure SQL DB General Purpose	$99\% < U_t \leq 99.99\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%
GCP	Google Cloud SQL	$99\% < U_t \leq 99.95\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%

Fonte: (AWS, 2021b)(AZURE, 2020a)(GCP, 2021b)

5.4.1.5 Base de dados não-relacional

No caso das bases não-relacionais, o melhor nível de serviço é oferecido pela AWS. Não é possível estabelecer uma ordem para os outros dois provedores, em decorrência da disjunção entre os intervalos de disponibilidade oferecidos para cada faixa de desconto.

Quadro 7 – SLAs oferecidos pelos provedores analisados para base de dados não-relacional na nuvem.

Provedor	**Serviço**	***Disponibilidade*(U_t)**	**% estornada em crédito**
AWS	Amazon DynamoDB Global Tables	$99\% \leq U_t < 99.999\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%
AWS	Amazon DynamoDB Standard	$99\% \leq U_t < 99.99\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	100%
Azure	CosmosDB Provisioned	$99\% < U_t \leq 99.99\%$	10%
		$U_t \leq 99\%$	25%
Azure	CosmosDB Serverless	$99\% < U_t \leq 99.9\%$	10%
		$U_t \leq 99\%$	25%
GCP	Google Cloud Multi-Region Datastore	$99\% < U_t \leq 99.95\%$	10%
		$95\% < U_t \leq 99\%$	25%
		$U_t \leq 95\%$	50%
GCP	Google Cloud Regional Datastore	$98\% < U_t \leq 99.95\%$	10%
		$95\% < U_t \leq 98\%$	25%
		$U_t \leq 95\%$	50%

Fonte: (AWS, 2021e)(AZURE, 2021b)(GCP, 2021e)

5.4.2 Outages

Mesmo com infraestruturas de hardware e software resilientes e redundantes, catástrofes e falhas podem acontecer, de modo que um ou mais serviços fiquem indisponíveis temporariamente para uma determinada região ou parte dela.

Outage é um termo originalmente usado para caracterizar interrupções no sistema de distribuição de energia elétrica que foi reutilizado, por analogia, no contexto de sistema distribuídos, caracterizando um período de tempo de indisponibilidade em algum serviço de computação, ou de um sistema como um todo(DICTIONARY, 2021)(GCP, 2021o).

Outages são extremamente projudiciais para os provedores de serviços na nuvem. Eles reduzem a disponibilidade do provedor, reduzem drasticamente os lucros (conforme apresentado na Subseção 5.4.1, em decorrência das multas de infrações dos SLAs) e não isentam os provedores de sofrerem processos legais (principalmente nos casos em que a catástrofe causa perda permanente de dados do cliente)(SULLIVAN, 2017).

Distinguir casos em que a indisponibilidade do serviço é causada por problemas locais e regionais de conexão, degradações parciais do serviço e outages pode não ser uma tarefa simples. Para esse propósito, os provedores fornecem geralmente um *dashboard* onde os consumidores podem verificar a disponibilidade de cada um dos serviços, em cada uma das regiões onde eles são oferecidos, com base em dados obtidos a partir de diferentes pontos geográficos ao redor do mundo (AWS, 2021o)(AZURE, 2021e)(GCP, 2021m).

Embora os provedores forneçam um histórico dos outages, em alguns casos, são publicados apenas os incidentes mais relevantes e/ou excluindo incidentes em que há apenas a degradação parcial de um serviço (AWS, 2021k)(AZURE, 2021f)(GCP, 2021n). A subjetividade desses critérios, combinada com os conflitos de interesses dos provedores, torna as fontes oficiais inadequadas para estabelecimento de uma análise comparativa. Para esse propósito, considerou-se o histórico fornecido por uma aplicação de terceiros.

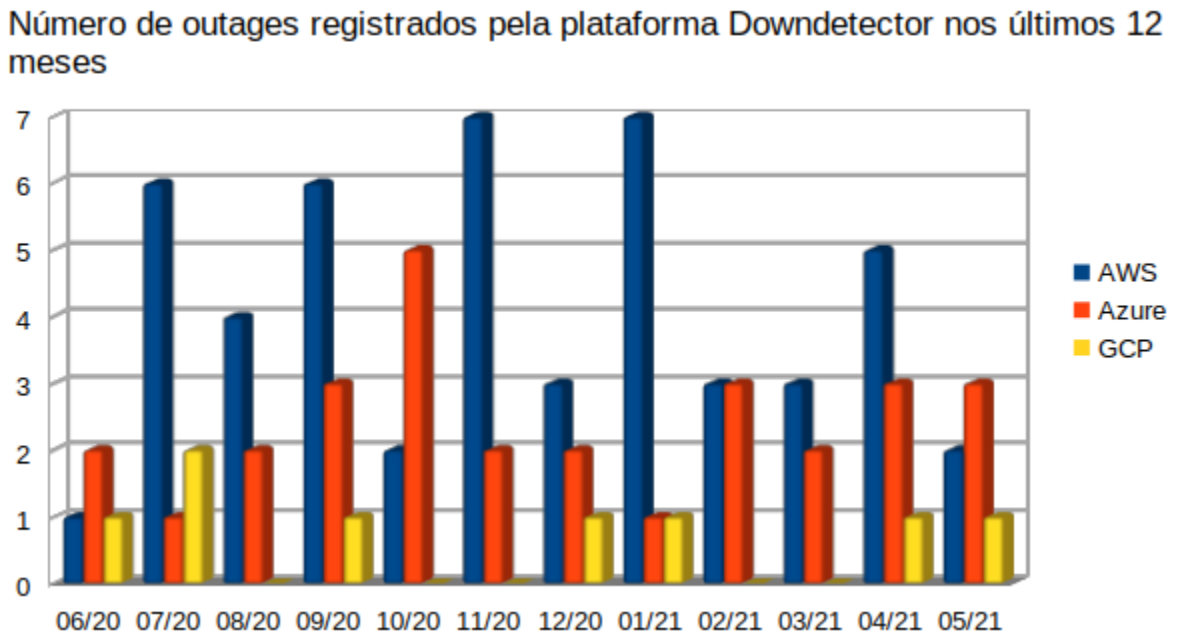
A aplicação Downdetector[3] oferece uma plataforma para monitoramento de *outages* em diversos serviços, baseado em mais de 22 milhões de relatórios mensais, fornecidos explicitamente por usuários ou coletados em redes sociais. Após a agregação dos dados de diversas fontes, executa-se um algorítmo de relevância com os resultados filtrados, eliminando relatórios de indisponibilidade local.

Considerando as informações disponibilizadas pelo Downdetector, foi possível contabilizar o número de casos reportados nos últimos 12 meses para cada um dos provedores. O resultado da contagem é representado no gráfico da Figura 18.

Uma análise rápida dos dados dispostos na Figura 18 permite ao leitor identificar que o maior número de incidentes acumulados é registrado para AWS, seguida da Azure e GCP, nessa ordem.

[3] Ferramenta disponível para acesso público em https://downdetector.com/

Figura 18 – Contagem dos incidentes reportados mensalmente por usuários, com validação e distribuição através da plataforma Downdetector.



Fonte: (DOWNDETECTOR, 2021)

.

5.5 BENCHMARKS DE DESEMPENHO DA INFRAESTRUTURA DE REDE

Um aspecto extremamente importante a ser considerado quando se faz uso de serviços de computação na nuvem é o desempenho da comunicação com o provedor. Um dos fatores que mais contribuem para a aferição do desempenho são as condições de rede e conectividade. A Internet é, por sua vez, a principal rede a partir da qual consumidores conectam-se a provedores públicos. Por esse motivo, nessa seção realiza-se uma análise de condições de rede das quais dispõem os provedores.

Os dados e as análises apresentados nessa seção foram extraídos do relatório anual conduzido pela instituição ThousandEyes para avaliar o desempenho e a qualidade da infraestrutura de comunicação oferecidos pelos principais provedores de serviços na nuvem (THOUSANDEYES, 2019-2020).

Os eventos da análise foram coletados em ambiente de produção, em múltiplas regiões de serviço, para os cinco provedores que são alvo da análise do relatório (AWS, Azure, GCP, IBM Cloud e Alibaba Cloud), durante o período de quatro semanas. O número de eventos coletados nesse período foi da ordem de 320 milhões.

Os testes de rede utilizam pacotes TCP para coletar dados de rede a cada *hop* (passo entre um elementos de roteamento) percorrido no caminho ao destinatário. As principais medições se dão por meio dos parâmetros de perda de pacotes, latência e *jitter* (variação

de latência), obtidos e medidos bi-direcionalmente entre as entidades observadas.

Nesse trabalho, são consideradas as seguintes métricas:

- **Medições de usuário final:** métricas de desempenho de rede obtidas a partir do consumidor de serviços da nuvem. As métricas consideram os serviços provisionado em múltiplas regiões de nuvem, sendo utilizados a partir de múltiplas localizações geográficas.
- **Medições de comunicação inter-regiões e inter-AZs:** métricas de desempenho entre as diversas regiões e AZs do mesmo provedor.

Nas subseções a seguir, detalha-se os experimentos realizados e indexa-se os resultados mais relevantes obtidos na análise de cada tipo de medição.

5.5.1 Medições de usuários finais

As medições de usuário final são importantes, pois medem os parâmetros de qualidade da comunicação entre o provedor e o consumidor. Quanto maior o tráfego de uma aplicação entre consumidor e provedor, maior a influência desses parâmetros no desempenho do sistema.

5.5.1.1 Metodologia

Métricas de desempenho de rede e conectividade foram coletadas a cada 10 minutos a partir de 98 pontos de medição, distribuídos em diversos *data centers* em 95 regiões de nuvem dos cinco provedores analisados pelo estudo. Os pontos de medição estão alojados em ISPs de Tier 2 e 3[4], distribuídos uniformemente ao longo do globo.

As localizações dos 98 pontos utilizados para medição de usuário final estão representadas no mapa da Figura 19. Embora testes tenham sido realizados em 95 regiões da nuvem, apenas um subconjunto dessas regiões é alvo da análise. A Figura 20 contém a lista das regiões que são alvo da análise, bem como uma relação de equivalência geográfica para as regiões de nuvem dos provedores avaliados.

5.5.1.2 Resultados

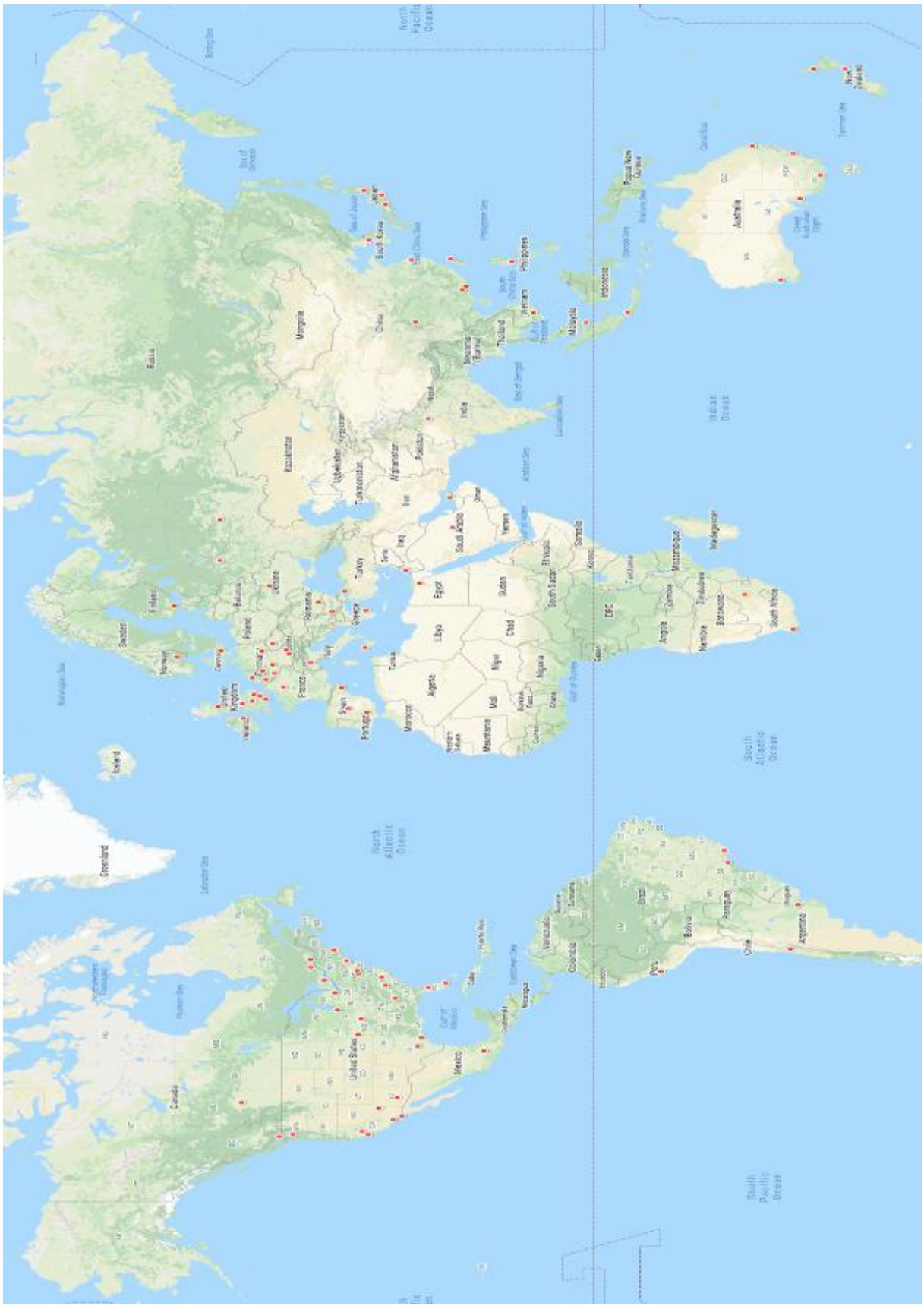
- **Evidências:** Os três provedores apresentaram desempenho de rede robusto e equivalente em regiões da América do Norte e Europa (salvo algumas exceções como regiões de nuvem em Singapura e Alemanha). A plataforma GCP apresentou uma

[4] Tiers são categorias atribuídas de acordo com a posição do ISP na hierarquia de infraestrutura global da Internet. Essas categorias são classificações não-regulamentadas e sem uma definição formal, mas, para efeito de compreensão, pode-se considerar uma tier 2 como uma rede com conexão direta à Internet e, uma tier 3 como uma rede com conexão indireta à Internet, ou seja, que acessa a Internet através de ISPs tier 2, por exemplo.

latência de rede de 2.5 a 3 vezes maior que as concorrentes AWS e Azure para regiões de nuvem na Índia, a partir da Europa. A maior diferença entre latências ocorreu utilizando as regiões de nuvem localizadas em São Paulo, a partir do Rio de Janeiro. Nesse caso, a latência da GCP foi 6 vezes maior que as concorrentes, como pode-se verificar na Figura 21.

- **Conclusão:** Os provedores oferecem desempenho satisfatório e condizente com as condições de infraestrutura locais de cada região política. Exceções foram encontradas na Ásia e Brasil, desfavorecendo a plataforma GCP, especialmente no critério de latência.

Figura 19 – Mapa com cada uma das 98 localizações geográficas dos pontos de medições de usuário final, representadas pelos marcadores vermelhos.



Fonte: Dados fonecidos por (THOUSANDEYES, 2019-2020)

.

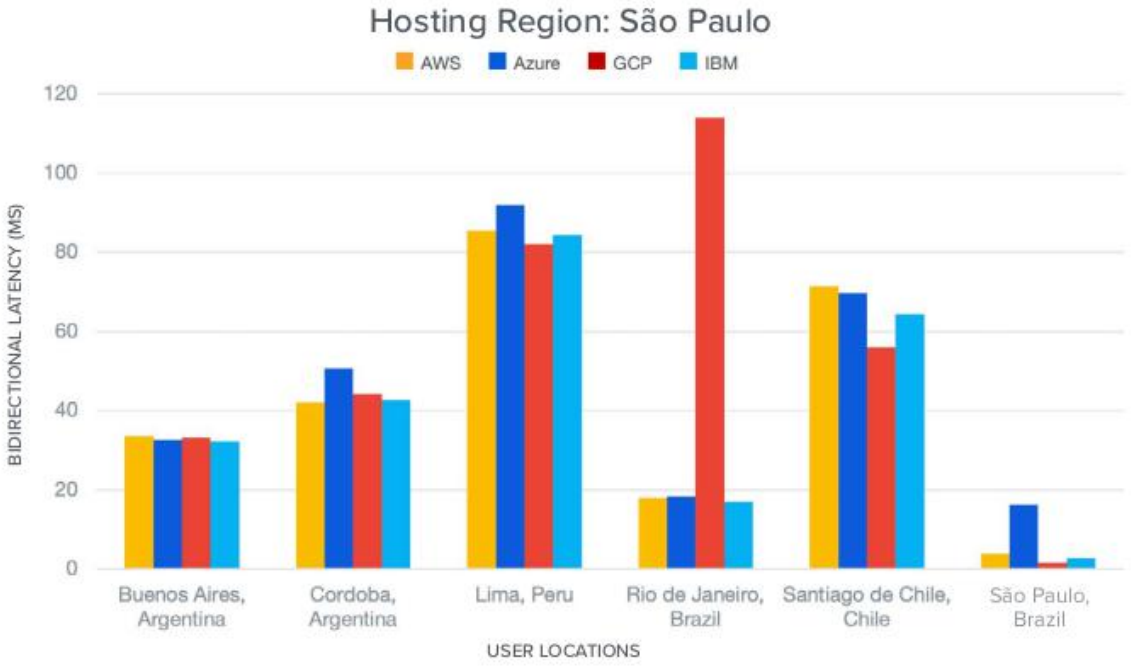
Figura 20 – Tabela com sub-conjunto de regiões consideradas na análise e equivalência geográfica entre regiões da nuvem de diferentes provedores.

	AMAZON WEB SERVICES	MICROSOFT AZURE	GOOGLE CLOUD PLATFORM
United States East	**us-east-1** Ashburn, VA	**East US** Ashburn, VA	**us-east4** Ashburn, VA
United States West	**us-west-1** San Jose, CA	**West US** Santa Clara, CA	**us-west2** Los Angeles, CA
United States Central	**us-east-2** Columbus, OH	**Central US** Des Moines, IA	**us-central1** Council Bluffs, IA
Canada	**ca-central-1** Montreal, Canada	**Canada East** Quebec City, Canada	**northamerica-northeast1** Montreal, Canada
South America	**sa-east-1** São Paulo, Brazil	**Brazil South** São Paulo, Brazil	**southamerica-east1** São Paulo, Brazil
Europe – London / Cardiff	**eu-west-2** London, UK	**UK West** Cardiff, Uk	**europe-west2** London, UK
Europe – Paris	**eu-west-3** Paris, France	**France Central** Paris, France	NA
Europe – Frankfurt	**eu-central-1** Frankfurt, Germany	NA	**europe-west3** Franfurt, Germany
Europe – Benelux	NA	**West Europe** Amsterdam, Netherlands	**europe-west4** Eemshaven, Netherlands
Asia – Singapore	**ap-southeast-1** Singapore	**Southeast Asia** Singapore	**asia-southeast1** Singapore
Asia – India	**ap-south-1** Mumbai, India	**West India** Mumbai, India	**asia-south1** Mumbai, India
Apac – Tokyo	**ap-northeast-1** Tokyo, Japan	**Japan East** Tokyo, Japan	**asia-northeast1** Tokyo, Japan
Apac – Australia	**ap-southeast-2** Sydney, Australia	**Australia East** Sydney, Australia	**australia-southeast1** Sydney, Australia

Fonte: (THOUSANDEYES, 2019-2020)

.

Figura 21 – Gráfico da latência bidirecional de comunicação provedor-consumidor por localidade geográfica de consumidor. A maior discrepância entre os valores apresentados acontece para consumidores situados na cidade do Rio de Janeiro, utilizando serviços da GCP provisionados na região de nuvem de São Paulo.



Fonte: (THOUSANDEYES, 2019-2020)

.

5.5.2 Medições de comunicação inter-regiões e inter-AZs

As medições inter-regiões medem a eficiência e qualidade da comunicação entre diferentes regiões de um mesmo provedor, essa característica é especialmente importante para esquemas globais de redundância e sistemas distribuídos em diferentes regiões do mundo. A comunicação entre AZs é especialmente importante para esquemas de replicação local, isto é, dentro de uma mesma região de nuvem.

5.5.2.1 Metodologia

Métricas referentes à comunicação entre as diferentes regiões de nuvem e AZs do mesmo provedor. As métrica são coletadas a cada 10 minutos partir de 6 regiões da AWS (us-east-1, us-west-1, sa-east-1, eu-west-2, eu-west-3 ap-south-1), 6 regiões da GCP (us-east4, us-west1, europe-west-2, asia-south1, asia-southeast1, southamerica-east1) e 4 regiões da Azure (East US, Central US, North Europe e France Central).

Embora haja uma pequena variação de distância entre duas regiões de um mesmo provedor, quando comparado a outro provedor, entende-se que essas variações não são significantes em magnitude a ponto de afetarem a ordem do resultado da comparação. Para cada um dos provedores analisados, foram considerados 15 pares de comunicação inter-regionais.

Dentro dos mesmos critérios, as medidas de desempenho de comunição entre AZs de uma mesma região foram realizadas para um subconjunto de regiões de cada um dos provedores analisados.

5.5.2.2 Resultados

Resultado 1:

- **Evidência:** O tráfego de rede inter-região permanece na infraestrutura privada de cada provedor.
- **Conclusão:** A utilização de infraestruturas privadas para tráfego inter-região torna essa comunicação menos suscetível a variações na qualidade de serviço e congestionamentos.

Resultado 2:

- **Evidência:** Para a plataforma GCP, o tráfego entre a região de nuvem do Sul Asiático (Mumbai) e algumas regiões de nuvem da Europa (Frankfurt, Londres, Bélgica e Holanda) apresentaram uma latência de rede 30% maior que os concorrentes.
- **Conclusão:** O tempo de latência de comunicação entre os pares de regiões selecionados foi homogêneo entre as regiões. Exceções podem ocorrer entre pares de regiões específicos.

Resultado 3:

- **Evidência:** A latência de comunicação das AZs dentro de cada uma das regiões analisadas fica entre 0.5 [ms] e 2.5 [ms]. A média para cada um do provedores foi: AWS (810 [μs]), Azure (740 [μs]) e GCP (520 [μs]).
- **Conclusão:** Considerando a média das medições inter-AZs de cada um dos provedores analisados, nota-se que todos apresentam resultados comparáveis, entretanto, é possível estabelecer a ordem decrescente em velocidade: GCP, Azure e AWS.

5.6 TARIFAÇÃO

As informações desta seção foram obtidas com base nos documentos oficiais dos provedores (AWS, 2021l)(AZURE, 2021d)(GCP, 2021k). Esta seção descreve alguns dos fatores mais relevantes no cálculo de tarifação, e compara o valor de custo para casos de uso comuns.

Alguns fatores que influenciam o custo de tarifação:

- Tipo de serviço (computação, armazenamento, etc.);
- Capacidade, quantidade e frequência de utilização dos recursos alocados;
- Modelo de precificação: sob demanda, instâncias reservadas ou planos de descontos;
- Prazo de reserva, para instâncias reservadas ou planos de descontos;
- Descontos de adiantamento de pagamento;
- Licenciamento de softwares proprietários;
- Regras contratuais *ad-hoc*, que não são objetos de análise desse trabalho.

Sobre os modelos de precificação, vale a definição dos modelos mais comuns entre os provedores públicos:

- **Sob-demanda:** Consiste na cobrança de acordo com a quantidade de unidades de tempo em que há uso do serviço. Quanto menor a unidade de tempo, mais fiel a cobrança será em relação à utilização real do recurso. Tipicamente, os provedores públicos oferecem unidades de tempo na ordem de minutos ou segundos.
- **Instâncias reservadas:** Planos de desconto baseados em comprometimento de volume de consumo por período de tempo. A redução do custo pode variar de acordo com a flexibilidade de utilização dos recursos cobertos pelo plano. Esse tipo de desconto é aplicado geralmente sobre contratos de reserva com duração da ordem de anos (usualmente, 1 e 3 anos).

Tipicamente, máquinas virtuais podem apresentar variantes de acordo com a otimização dos seus recursos, de modo que se tornem mais ou menos apropriadas para um tipo de tarefa e, naturalmente, possuam valores de tarifação distintos. São os principais tipos de instâncias de máquinas virtuais (AWS, 2021m):

- **Propósito geral:** Instâncias com recursos de processamento, memória e rede balanceados. Indicadas para aplicações em geral, com consumo de recursos balanceado, como servidores Web.
- **Computação otimizada:** Instâncias com uso intenso de recursos de processamento, como computação de alto desempenho, modelagem científica, *workers* com alta carga de trabalhos em lote e servidores de alto desempenho. Indicadas para aplicações *CPU-bound* em geral.
- **Memória otimizada:** Instâncias otimizadas para entregar bom desempenho para cargas de trabalho que envolvem o processamento de grandes conjuntos de dados, em memória. Indicada para aplicações com uso intenso de recursos de memória.
- **Computação acelerada:** Instâncias equipadas com aceleradores de hardware, ou co-processadores, para otimizar o processamento envolvendo operações com ponto flutuante e cálculos matriciais (frequentes em computação gráfica). Indicada para aplicações que podem ter parte de sua carga de trabalho paralelizada, através do uso de aceleração de hardware.
- **Armazenamento otimizado:** Instâncias otimizadas para a realização de operações de leitura e escrita do armazenamento local com baixa latência. Indicadas para aplicações que realizam dezenas de milhares de operações de entrada e saída por segundo (IOPS).

5.6.1 Metodologia

A variedade de recursos, configurações de recursos e modelos de tarifação inviabiliza uma análise completa do esquema de tarifação dos provedores analisados.

Por essa razão, nesta seção apresenta-se uma comparação restrita aos valores de cobrança de três dos serviços mais fundamentais da nuvem: computação, armazenamento e tráfego de rede, de forma conjunta, sob as seguintes configurações:

- **Região:** São Paulo, BR (AWS: sa-east-1, Azure: Brazil South, GCP: southamerica-east1);
- **Tipo de instância:** Propósito geral, não-preemptiva[5];

[5] É possível obter tarifas menores utilizando instâncias preemptivas, isto é, instâncias que podem ter sua execução interrompida a qualquer momento para atendimento de outro *tenant*. Esse tipo de comportamento pode ser tolarável em algumas aplicações, como processamento assíncrono de dados.

- **Número de VCPUs:** 4;
- **Memória RAM:** 16 [GB];
- **Sistema operacional:** Linux Ubuntu;
- **Armazenamento persistente:** Armazenamento HDD padrão de 1 [TB] (AWS: sc1, Azure: Standard HDD, GCP: Standard Storage);
- **Tráfego de dados egressante para a Internet:** 64 [GB/mes];
- **Porcentagem de utilização diária sob-demanda:** 1% (730 horas por mês);
- **Adiantamento de pagamento:** 0 [USD];

Para os recursos descritos, calcula-se a estimativa para consumo sob-demanda e com instâncias reservadas de 1 e de 3 anos. Os valores apresentados na análise foram obtidos usando as calculadoras oficiais de estimativa de custo, oferecidas por cada um provedores analisados (AWS, 2021n)(AZURE, 2021g)(GCP, 2021l).

5.6.2 Resultados

Os resultados obtidos com a simulação de custo do conjunto de recursos provisionados, de acordo com a descrição da metodologia, permitiu classificar os provedores em uma ordem crescente de custo: GCP (165,34 USD), AWS (167,80 USD) e Azure (237,69 USD) (Figura 22).

A plataforma Azure não só obteve o maior custo total, como também os maiores valores para os serviços de computação e armazenamento, isoladamente. No caso do custo associado ao consumo de banda de rede, em decorrência do tráfego de dados egressantes, obteve-se a ordem crescente de custo: Azure, GCP, AWS. Entretanto, a magnitude dessa cobrança contribui apenas com uma pequena parcela do custo total.

Listam-se algumas observações relativas ao custo dos serviços:

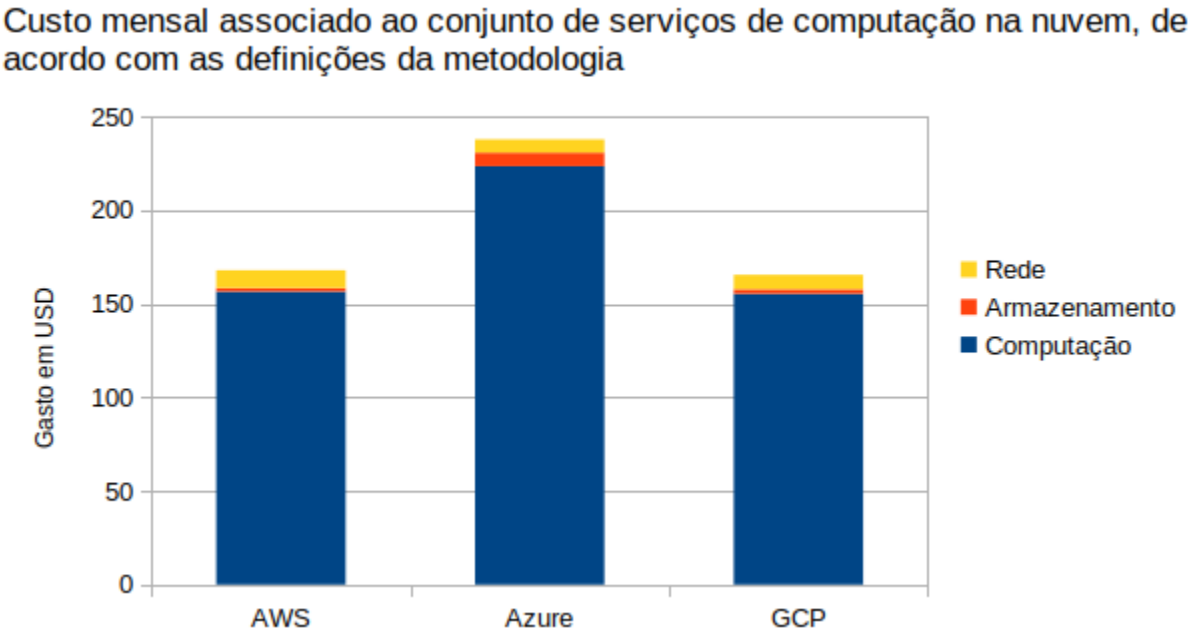
- As plataformas AWS e GCP possuem custos totais equivalentes;
- A plataforma Azure apresentou custo total mais de 40 % maior que as concorrentes;
- As plataformas AWS e GCP possuem custos de serviços de computação equivalentes;
- A plataforma Azure apresentou custo de serviços de computação (223,38 USD) mais de 42% maior que as concorrentes;
- A plataforma GCP possui custo de armazenamento de 2,36 USD mensais, valor 23 % maior que a melhor tarifa, oferecida pela AWS (1,84 USD);

- A plataforma Azure possui custo de armazenamento mais de 200 % maior que o colocado acima (GCP);
- Os custos associados ao consumo de banda por dados egressantes possui tarifas compatíveis entre os provedores analisados.

O custo associado ao consumo de recursos computacionais decresce de forma logarítmica e inversamente proporcional ao comprometimento de uso (Figura 23). Por esse motivo, instâncias reservadas são opções extremamente viáveis para cargas de trabalho com comportamento estável.

Dentre os provedores analisados, as melhores porcentagens de desconto para instâncias reservadas foram encontradas para a plataforma Azure (com até 62% de desconto), seguida das plataformas AWS (com até 57% de desconto) e GCP (com até 55% de desconto).

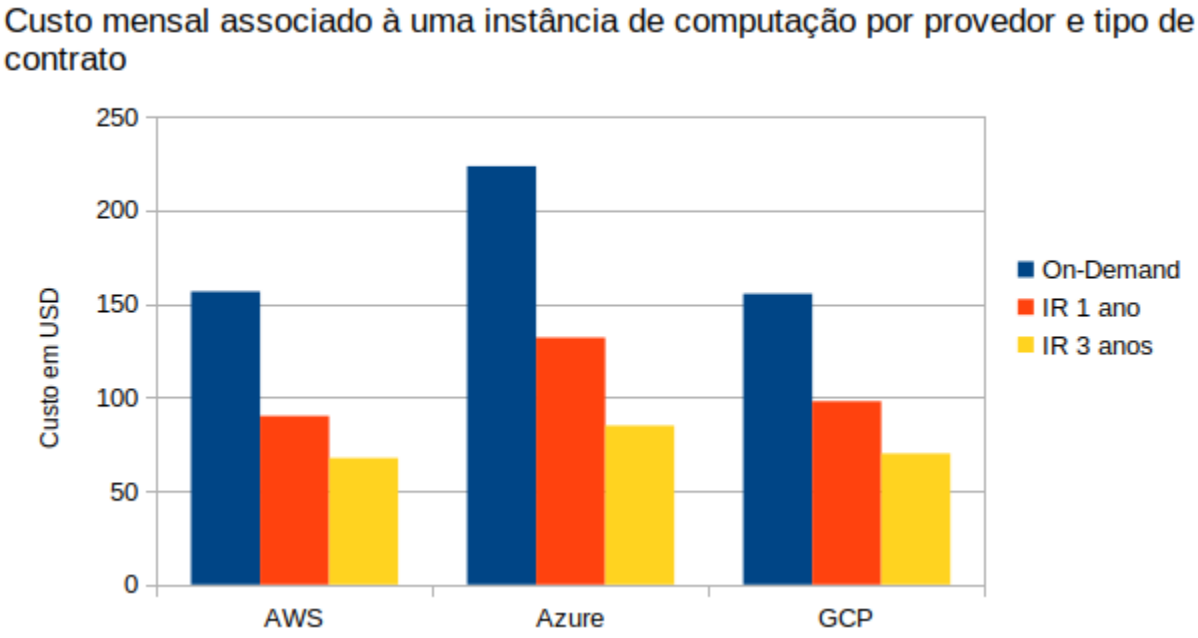
Figura 22 – Gráfico com os custos associados ao provisionamento de máquinas virtuais, armazenamento e banda de rede, de acordo com a especificação determinada pela metodologia.



Fontes: (AWS, 2021n)(AZURE, 2021g)(GCP, 2021l)

.

Figura 23 – Gráfico com os custos associados ao provisionamento de máquinas virtuais, comparando cada um dos provedores, para consumo sob-demanda e instâncias reservadas (IR) por 1 ou 3 anos.



Fontes: (AWS, 2021n)(AZURE, 2021g)(GCP, 2021l)

.

5.7 CERTIFICAÇÕES

Grande parte das aplicações empresariais demandam requisitos especiais de conformidade (*compliance*). Para grande parte dos casos de uso, o cumprimento de um conjunto de especificações pode ser um fator mandatório na escolha do provedor.

Por esse motivo, os provedores públicos buscam atender o maior número possível de certificações relevantes para o consumidor. As certificações quase sempre contam com regras e parâmetros para garantia da segurança e padrões de qualidade.

De 38 certificações analisadas, constatou-se que a Azure cumpre 34, a AWS cumpre 33 e a GCP cumpre 24, estabelecendo essa ordem como resultado de um comparativo quantitativo, conforme pode-se verificar na Tabela 8.

Quadro 8 – Relação de atendimento de certificações por provedores.

#	**AWS**	**GCP**	**Azure**
CSA	Sim	Sim	Sim
ISO 9001	Sim	Não	Sim
ISO 27001	Sim	Sim	Sim
EU Model Clauses	Sim	Sim	Sim
PCI DSS Level 1	Sim	Sim	Sim
ISO 27017	Sim	Sim	Sim
ISO 22301	Não	Não	Sim
ISO 27018	Sim	Sim	Sim
ISO 31000	Não	Não	Não
SOC 1	Sim	Sim	Sim
SOC 3	Sim	Sim	Sim
HITRUST	Não	Sim	Sim
GDPR	Sim	Sim	Sim
CJIS	Sim	Não	Sim
DoD SRG	Sim	Não	Sim
FedRAMP	Sim	Sim	Sim
FIPS	Sim	Sim	Sim
FERPA	Sim	Sim	Sim
FFIEC	Sim	Não	Sim
HIPAA	Sim	Sim	Sim
ITAR	Sim	Não	Sim
GxP	Sim	Não	Sim
MPAA	Sim	Sim	Sim
NIST	Sim	Sim	Sim
FISMA	Sim	Não	Sim
EU-US Privacy Shield	Não	Sim	Sim
FISC	Sim	Sim	Sim
IRAP	Sim	Sim	Sim
K-ISMS	Sim	Não	Não
MTCS Tier3	Sim	Sim	Sim
My Number Act	Sim	Sim	Sim
C5	Sim	Sim	Sim
Cyber Essentials Plus	Sim	Sim	Sim
ENS High	Sim	Sim	Sim
G - Cloud	Sim	Não	Sim
IT-Grundschutz	Sim	Não	Sim
TISAX	Sim	Não	Não
UKCSP	Não	Não	Não

Fonte: (CLOUDCOMPARISONTOOL, 2021)

5.8 DISCUSSÃO

Como foi possível observar ao longo das seções deste capítulo, a caracterização dos provedores da nuvem varia de acordo com diversos fatores (como serviço e localidade). Desse modo, discute-se nessa seção, superficialmente, uma proposta de ordenação parcial dos provedores, de acordo com cada aspecto analisado, isoladamente.

Considerando os dados apresentados na Seção 5.1, constata-se o crescimento da relevância dos serviços de computação na nuvem, mediante ao crescimento econômico do setor.

A análise dos gastos de usuários, associados ao consumo de serviços na nuvem, identifica como três princpais provedores: AWS, Azure e GCP, nessa ordem. Considerando os níveis de adoção dos provedores dentro do contexto empresarial, tem-se a ordem mantida, porém com uma grande aproximação entre a AWS e os concorrentes Azure e GCP.

Análises adicionais permitem obter alguns resultados gerais de utilização da nuvem:

- O uso intenso de nuvens híbridas e multi-clouds indica que 92% dos consumidores fazem uso simultâneo de mais de um provedor;
- O modelo de serviço mais popular na nuvem é SaaS, seguido de IaaS e PaaS, nessa ordem.

Considerando os dados apresentados na Seção 5.3, entende-se que não é viável uma comparação direta do número de regiões em decorrência da divergência de definição utilizada pelos provedores. Como não foi possível identificar o número exato de AZs da Azure, realiza-se a comparação desse aspecto baseada no número de territórios atendidos. Sob esse aspecto, temos a ordem como: AWS (245), GCP (200+) e Azure (140+).

Considerando os dados apresentados na Seção 5.4, obtem-se duas classificações, uma com relação aos SLAs oferecidos, que forçam financeiramente os provedores a manterem níveis aceitáveis de disponibilidade, e outra com relação aos registros de incidentes de indisponibilidade ou degradação dos serviços.

Da perspectiva dos SLAs, nota-se que há variação na ordenação dos provedores dentre os serviços analisados e, em casos raros, a impossibilidade de estabelecimento de uma ordenação total. Das 5 categorias de serviço analisadas, 3 obtiveram o melhor acordo com GCP, 1 com Azure e 1 com AWS. Nessas condições, considera-se o GCP a plataforma que oferece os melhores SLAs.

Com relação ao número de incidentes registrados, temos a ordem crescente: GCP (8), Azure (29) e AWS (49). Entende-se que a diferença de número de usuários e carga de trabalho entre as plataformas podem ter um efeito sobre as contagens apresentadas, fazendo necessária uma normalização sobre os dados para aumento da precisão dos resultados. Entretanto, em decorrência das grandes distâncias entre os somatórios de incidentes para

cada provedor, ainda pode-se classificar a plataforma GCP como a mais resiliente dentre as analisadas.

Considerando os dados apresentados na Seção 5.5, nota-se que, de todas as análises do capítulo, essa é a que possui maior grau de regionalidade. Os testes de desempenho de usuário final identificaram discrepância entre a latência dos provedores para poucas regiões. Entre as medidas, destaca-se a latência da plataforma GCP para a região de São Paulo, que foi mais de 5 vezes maior que os valores apresentados pelos concorrentes. A análise de tráfego inter-região e inter-AZs permitiu identificar maior estabilidade nas plataformas GCP e Azure, em decorrência do uso mais intenso de infraestruturas de rede privadas para esse tipo de comunicação.

Considerando os dados apresentados na Seção 5.6, nota-se que, devido a complexidade e multiplicade de modelos, um estudo completo das regras de tarifação é inviável para o escopo deste trabalho.

Desse modo, a partir da simulação de um caso de uso foi possível fazer uma análise *ad-hoc* da tarifação nos diferentes provedores e o decaimento dos preços mediante ao uso de planos de desconto para instâncias de máquinas virtuais (excelentes para execução de cargas de trabalho com demanda conhecida a priori). A análise permitiu identificar um custo inferior e aproximadamente igual para as plataformas GCP e AWS. A plataforma Azure apresentou custo total mais de 40% maior que as concorrentes. No caso dos serviços de armazenamento, a plataforma Azure apresentou valor 200% supeior à oferta da GCP.

Sumarizando, classifica-se AWS e GCP como as plataformas com menor custo para o consumidor para o caso de uso proposto pela Seção 5.6.

Considerando os dados apresentados na Seção 5.7, constata-se que a ordem decrescente obtida, comparando o número das certificações é: Azure (34), AWS (33), GCP (24).

6 CONCLUSÃO

Esse trabalho provê ao leitor uma fonte unificada para compreender o modelo de computação na nuvem em *lato sensu*, a partir de um conjunto sólido de fontes bibliográficas. Inicialmente, apresenta-se cada uma das tecnologias consideradas fundamentais para a realização de computação na nuvem: redes de computadores (representadas pela Internet), computação em grade e virtualização. Na sequência, define-se formalmente o modelo de computação na nuvem, seguido de sua história sintetizada. Finalmente, materializa-se a análise por meio da comparação entre os provedores públicos mais relevantes de computação na nuvem: AWS, GCP e Azure.

Durante os quase 40 anos entre as primeiras proposições de computação utilitária e a consolidação do modelo de computação na nuvem, verifica-se o desenvolvimento de cada uma das tecnologias consideradas fundamentais.

A combinação das tecnologias de computação em grade e virtualização, tornaram possível o surgimento das primeiras nuvens privadas, no fim do Século 20. No início do Século 21, o estabelecimento e disseminação da infraestrutura pública de Internet, em diversas regiões do mundo, viabilizou a implantação do modelo público de serviço na nuvem.

Em 2006, a plataforma AWS foi o primeiro provedor a oferecer serviços no modelo IaaS. Desde então, o modelo de computação na nuvem tem se tornado cada vez mais promissor e popular, representando atualmente uma parcela considerável dos gastos de usuários finais com computação em geral.

De acordo com as definições providas por (MELL; GRANCE et al., 2011) e fontes bibliográficas complementares, a nuvem pode ser fundamentalmente definida a partir de cinco características essenciais, três modelos de serviço e quatro modelos de *deployment*.

São as cinco características essenciais:

i) *Self-service* sob-demanda; ii) Amplo acesso à rede de computadores; iii) *Pooling* de recursos; iv) Rápida elasticidade; v) Serviços com métricas[1].

São caracterizados os três modelos de serviço:

i) Infraestrutura como um Serviço (IaaS); ii) Plataforma como um Serviço (PaaS); iii) Software como um Serviço (SaaS).

São caracterizados os quatro modelos de *deployment*:

i) Público; ii) Privado; iii) Comunitário; iv) Híbrido.

Na comparação de provedores públicos AWS, Azure e GCP são identificados como os provedores de maior adoção e participação econômica no cenário de computação na nuvem atual. Os provedores selecionados foram avaliados sobre os seguintes parâmetros: i)

[1] Essa característica está fortemente associada à tarifação e ao monitoramento de saúde e desempenho dos serviços providos.

mercado e popularidade; ii) oferta de serviços; iii) disposição geográfica de infraestrutura; iv) confiabilidade (caracterizada através da avaliação dos SLAs disponíveis e da contagem dos registros de *outages*, referentes a cada um dos provedores, individualmente, durante o período de 365 dias); v) *benchmarks* de desempenho da infraestrutura de rede; vi) tarifação (baseada na comparação de custo efetivo calculado para um caso de uso comum); e vii) certificações.

Embora não seja possível estabelecer uma ordem geral, no Capítulo 5, cada seção traz consigo uma discussão das vantagens e desvantagens de cada um dos provedores selecionados, em relação ao parâmetro analisado na seção.

A plataforma AWS permanece a líder do mercado em oferta de serviços de computação na nuvem. O número de países atendidos é o maior entre as concorrentes. O registro de um alto número de *outages* pode indicar instabilidade da plataforma para alguns serviços. A análise dos dados apresentados em (THOUSANDEYES, 2019-2020) permitiu identificar maior instabilidade para esse provedor, em comparação aos demais, justificada pela referência como resultado do uso mais intenso das infraestruturas públicas da Internet. Os custos menores, obtidos para o caso de uso analisado, favorecem o uso da plataforma. A variedade de certificações atendidas torna a plataforma viável para um número maior de aplicações e casos de uso.

A plataforma Azure segue na vice-liderança do mercado, com popularidade comparável à da líder AWS. A análise dos dados apresentados em (THOUSANDEYES, 2019-2020) permitiu identificar estabilidade e boa latência, ambos justificados pela referência como resultado da proximidade entre o consumidor e os pontos de ingresso na infraestrutura privada do provedor. Dos provedores analisados, a Azure apresentou o maior custo efetivo, com SLAs inferiores ao da concorrente GCP, para consumo dos recursos no caso de uso analisado. A variedade de certificações atendidas torna a plataforma viável para um número maior de aplicações e casos de uso.

A plataforma GCP ocupa o terceiro lugar na classificação dos provedores em relação ao mercado e à porcentagem de utilização. A grande porcentagem de usuários que fazem uso experimental da plataforma indica um possível crescimento, a curto prazo, no número de consumidores, com cargas de trabalho cada vez maiores. A análise dos dados apresentados em (THOUSANDEYES, 2019-2020) permitiu identificar estabilidade e boa latência, ambos justificados pela referência como resultado da proximidade entre o consumidor e os pontos de ingresso na infraestrutura privada do provedor. Em algumas regiões, a comunicação ponto-a-ponto apresentou demasiada latência para algumas localidades de consumo, o que pode tornar o provedor inviável de acordo com a localidade de origem e destino do fluxo de consumo dos serviços. A plataforma GCP apresentou os melhores SLAs, na maior parte dos casos analisados. Os custos menores, obtidos para o caso de uso analisado, favorecem o uso da plataforma. A carência de algumas certificações pode tornar a plataforma inviável para algumas aplicações e casos de uso.

A análise das estatísticas referentes ao uso dos provedores por parte das empresas indica um alto nível de concomitância no uso de diferentes provedores públicos, possivelmente em decorrência da variedade de oferta de serviços, com diferentes capacidades e preços. Cláusulas especificas de conformidade podem requerir que determinados dados ou cargas de trabalho permaneçam na infraestrutura *on-premise* do consumidor. Essas são algumas das razões pelas quais acreditamos que a ocorrência de *deployments* multinuvem e híbridos tenderá a se intensificar nos próximos anos.

Como proposta de trabalhos futuros, enuncia-se: i) implementação de um *framework* para medição e análise de dados de performance de um provedor; ii) análise de usabilidade das plataformas de nuvem oferecidas pelos provedores, do ponto de vista dos consumidores de serviços; iii) análise e atualização da arquitetura conceitual da nuvem proposta em (LIU et al., 2011); iv) elaboração de demonstrações matemáticas que comprovem a eficiência do modelo de computação na nuvem em relação aos demais modelos de computação; v) implementação e análise de ferramentas para manipulação de nuvens híbridas.

REFERÊNCIAS

ALEXANDROV, A. D. et al. Superweb: Towards a global web-based parallel computing infrastructure. In: IEEE. **Proceedings 11th International Parallel Processing Symposium**. [S.l.], 1997. p. 100–106.

Anderson, D. P. Boinc: a system for public-resource computing and storage. In: **Fifth IEEE/ACM International Workshop on Grid Computing**. [S.l.: s.n.], 2004. p. 4–10.

ANDERSON, D. P. et al. Seti@ home: an experiment in public-resource computing. **Communications of the ACM**, ACM New York, NY, USA, v. 45, n. 11, p. 56–61, 2002.

AVERY, P. et al. **An international virtual-data grid laboratory for data intensive science**. [S.l.], 2001.

AWS. **Announcing Amazon Elastic Compute Cloud (Amazon EC2) - beta**. 2006. Disponível em: https://aws.amazon.com/premiumsupport/technology/pes/.

AWS. **Announcing Amazon Elastic Compute Cloud (Amazon EC2) - beta**. 2006. Disponível em: https://aws.amazon.com/about-aws/whats-new/2006/08/24/announcing-amazon-elastic-compute-cloud-amazon-ec2---beta/.

AWS. 2021. Disponível em: https://aws.amazon.com/outposts/.

AWS. **Amazon Aurora Service Level Agreement**. 2021. Disponível em: https://aws.amazon.com/rds/aurora/sla/.

AWS. **Amazon CloudFront Service Level Agreement**. 2021. Disponível em: https://aws.amazon.com/cloudfront/sla/.

AWS. **Amazon Compute Service Level Agreement**. 2021. Disponível em: https://aws.amazon.com/compute/sla/.

AWS. **Amazon DynamoDB Service Level Agreement**. 2021. Disponível em: https://aws.amazon.com/dynamodb/sla/.

AWS. **Amazon S3 Service Level Agreement**. 2021. Disponível em: https://aws.amazon.com/s3/sla/.

AWS. **AWS Application Programming Interface (API)**. 2021. Disponível em: https://docs.aws.amazon.com/general/latest/gr/aws-apis.html.

AWS. **AWS Command Line Interface (CLI)**. 2021. Disponível em: https://aws.amazon.com/cli/.

AWS. **AWS Console**. 2021. Disponível em: https://console.aws.amazon.com/.

AWS. **AWS On-demand pricing**. 2021. Disponível em: https://aws.amazon.com/ec2/pricing/on-demand/.

AWS. **AWS Post-Event Summaries**. 2021. Disponível em: https://aws.amazon.com/premiumsupport/technology/pes/.

AWS. **AWS pricing**. 2021. Disponível em: https://aws.amazon.com/pricing/.

AWS. **AWS pricing**. 2021. Disponível em: https://aws.amazon.com/ec2/instance-types/.

AWS. **AWS Pricing Calculator**. 2021. Disponível em: https://calculator.aws/.

AWS. **AWS Service Health Dashboard**. 2021. Disponível em: https://status.aws.amazon.com/.

AWS. **AWS Service Level Agreements (SLAs)**. 2021. Disponível em: https://aws.amazon.com/legal/service-level-agreements/.

AWS. **How scaling plans work**. 2021. Disponível em: https://docs.aws.amazon.com/autoscaling/plans/userguide/how-it-works.html.

AWS. **Mapa da infraestrutura global da AWS**. 2021. Disponível em: https://aws.amazon.com/pt/about-aws/global-infrastructure/.

AWS. **Regions and Availability Zones**. 2021. Disponível em: https://aws.amazon.com/about-aws/global-infrastructure/regions_az/.

AZURE. **Azure Application Programming Interface (API)**. 2021. Disponível em: https://azure.microsoft.com/en-us/services/api-management/.

AZURE. **Azure Command Line Interface (CLI)**. 2021. Disponível em: https://docs.microsoft.com/en-us/cli/azure/install-azure-cli.

AZURE. **Azure Portal**. 2021. Disponível em: https://portal.azure.com/.

AZURE. **Azure pricing**. 2021. Disponível em: https://azure.microsoft.com/en-us/pricing/.

AZURE. **Azure status**. 2021. Disponível em: https://status.azure.com/en-us/status.

AZURE. **Azure status history**. 2021. Disponível em: https://status.azure.com/en-us/status/history/.

AZURE. **Pricing Calculator**. 2021. Disponível em: https://azure.microsoft.com/en-us/pricing/calculator/.

AZURE, M. **SLA for Content Delivery Network**. 2015. Disponível em: https://azure.microsoft.com/en-us/support/legal/sla/cdn/v1_0/.

AZURE, M. **SLA for Storage Accounts**. 2019. Disponível em: https://azure.microsoft.com/en-us/support/legal/sla/storage/v1_5/.

AZURE, M. **SLA for Azure SQL Database**. 2020. Disponível em: https://azure.microsoft.com/en-us/support/legal/sla/sql-database/v1_5/.

AZURE, M. **SLA for Virtual Machines**. 2020. Disponível em: https://azure.microsoft.com/en-us/support/legal/sla/virtual-machines/v1_9/.

AZURE, M. **Azure geographies**. 2021. Disponível em: https://azure.microsoft.com/en-us/global-infrastructure/geographies/.

AZURE, M. **SLA for Azure Cosmos DB**. 2021. Disponível em: https://azure.microsoft.com/en-us/support/legal/sla/cosmos-db/v1_4/.

AZURE, M. **SLA summary for Azure services**. 2021. Disponível em: https://azure.microsoft.com/en-us/support/legal/sla/summary/.

BARATLOO, A. et al. Charlotte: Metacomputing on the web. **Future Generation Computer Systems**, Elsevier, v. 15, n. 5-6, p. 559–570, 1999.

BERKELEY, U. **Choosing BOINC projects**. 2021. Disponível em: https://boinc.berkeley.edu/projects.php.

BERNERS-LEE, T. et al. The world-wide web. **Communications of the ACM**, ACM New York, NY, USA, v. 37, n. 8, p. 76–82, 1994.

BERNERS-LEE, T. J. **Information management: A proposal**. [S.l.], 1989.

BLOKDIJK, G.; MENKEN, I. **Cloud Computing - The Complete Cornerstone Guide to Cloud Computing Best Practices: Concepts, Terms, and Techniques for Successfully Planning, Implementing ... Cloud Computing Technology - Second Edition**. London, GBR: Emereo Pty Ltd, 2009. ISBN 1742441408.

BRECHT, T. et al. Paraweb: Towards world-wide supercomputing. In: **Proceedings of the 7th workshop on ACM SIGOPS European workshop: Systems support for worldwide applications**. [S.l.: s.n.], 1996. p. 181–188.

CAMIEL, N. The popcorn project: Distributed computation over the internet in java. In: **6th International World Wide Web Conference**. [S.l.: s.n.], 1997.

CAMPBELL, S.; JERONIMO, M. An introduction to virtualization. **Published in "Applied Virtualization", Intel**, p. 1–15, 2006.

CANNONICAL. 2021. Disponível em: https://ubuntu.com/download/desktop.

CARR, C. S.; CROCKER, S. D.; CERF, V. G. Host-host communication protocol in the arpa network. In: **Proceedings of the May 5-7, 1970, spring joint computer conference**. [S.l.: s.n.], 1970. p. 589–597.

CERF, V. G.; KAHN, R. E. A protocol for packet network intercommunication. **IEEE Trans. Commun.**, v. 22, n. 5, p. 637–648, 1974. Disponível em: https://doi.org/10.1109/TCOM.1974.1092259.

CLARK, D. D. et al. The aurora gigabit testbed. **Computer Networks and ISDN Systems**, Elsevier, v. 25, n. 6, p. 599–621, 1993.

CLOUDCOMPARISONTOOL. 2021. Disponível em: https://www.cloudcomparisontool.com/.

COMPARECLOUD.IN. **Public Cloud Services Comparison**. 2021. Disponível em: https://comparecloud.in/.

CURRY, J. **Introducing OpenStack, The OpenStack Blog**. 2010. Disponível em: https://web.archive.org/web/20171026111206/https://www.openstack.org/blog/2010/07/introducing-openstack/.

DARROW, B. **Exclusive: RightScale is first to resell, support Google Compute Engine**. 2013. Disponível em: http://gigaom.com/2013/02/25/exclusive-rightscale-is-first-to-resell-support-google-compute-engine/.

DICTIONARY, C. **Incidents and the Google Cloud Status Dashboard**. 2021. Disponível em: https://dictionary.cambridge.org/dictionary/english/outage.

DOCKER. **What is a Container?** 2021. Disponível em: https://www.docker.com/resources/what-container.

DOWNDETECTOR. 2021. Disponível em: https://downdetector.com/archive/.

EVANGELIST, A. 2012. Disponível em: https://apievangelist.com/2012/06/03/rise-of-mobile-backend-as-a-service-mbaas-api-stacks/.

FLEXERA. Flexera 2021 state of the cloud report. **Flexera State of the Cloud Report**, 2021.

FOSTER, I. What is the grid? a three point checklist. **GRID today**, v. 1, n. 6, p. 32–36, 2002.

FOSTER, I. Globus toolkit version 4: Software for service-oriented systems. **Journal of computer science and technology**, Springer, v. 21, n. 4, p. 513–520, 2006.

FOSTER, I.; KESSELMAN, C. Globus: A metacomputing infrastructure toolkit. **The International Journal of Supercomputer Applications and High Performance Computing**, Sage Publications Sage CA: Thousand Oaks, CA, v. 11, n. 2, p. 115–128, 1997.

FOSTER, I.; KESSELMAN, C. **The Grid 2: Blueprint for a new computing infrastructure**. [S.l.]: Elsevier, 2003.

FOSTER, I.; KESSELMAN, C. The history of the grid. In: **High Performance Computing: From Grids and Clouds to Exascale**. [S.l.]: IOS Press, 2011. p. 3–30.

FOWLER, M. 2018. Disponível em: https://martinfowler.com/articles/serverless.html#unpacking-faas.

GAGLIARDI, F. et al. European datagrid project: Experiences of deploying a large scale testbed for e-science applications. In: SPRINGER. **IFIP International Symposium on Computer Performance Modeling, Measurement and Evaluation**. [S.l.], 2002. p. 480–499.

GARFINKEL, S. **Architects of the information society: 35 years of the Laboratory for Computer Science at MIT**. [S.l.]: MIT press, 1999.

GCP. **Cloud CDN Service Level Agreement (SLA)**. 2021. Disponível em: https://cloud.google.com/cdn/sla.

GCP. **Cloud SQL Service Level Agreement (SLA)**. 2021. Disponível em: https://cloud.google.com/sql/sla.

GCP. **Cloud Storage Service Level Agreement (SLA)**. 2021. Disponível em: https://cloud.google.com/storage/sla.

GCP. **Compute Engine Service Level Agreement (SLA)**. 2021. Disponível em: https://cloud.google.com/compute/sla.

GCP. **Datastore Service Level Agreement (SLA)**. 2021. Disponível em: https://cloud.google.com/datastore/sla.

GCP. **gcloud command-line tool overview**. 2021. Disponível em: https://cloud.google.com/sdk/gcloud.

GCP. **GCP Application Programming Interface (API)**. 2021. Disponível em: https://cloud.google.com/apis/docs/overview.

GCP. **GCP Console**. 2021. Disponível em: https://console.cloud.google.com/.

GCP. **Google Cloud metrics**. 2021. Disponível em: https://cloud.google.com/monitoring/api/metrics_gcp.

GCP. **Google Cloud Platform Service Level Agreements**. 2021. Disponível em: https://cloud.google.com/terms/sla.

GCP. **Google Cloud pricing**. 2021. Disponível em: https://cloud.google.com/pricing.

GCP. **Google Cloud Pricing Calculator**. 2021. Disponível em: https://cloud.google.com/products/calculator.

GCP. **Google Cloud Status Dashboard**. 2021. Disponível em: https://status.cloud.google.com/.

GCP. **Google Cloud Status Dashboard**. 2021. Disponível em: https://status.cloud.google.com/summary.

GCP. **Incidents and the Google Cloud Status Dashboard**. 2021. Disponível em: https://cloud.google.com/support/docs/dashboard.

GCP. **Locais do Cloud**. 2021. Disponível em: https://aws.amazon.com/pt/about-aws/global-infrastructure/.

GOOGLE. **Google Workspace (antigamente G Suite)**. 2021. Disponível em: https://workspace.google.com/.

GOYAL, A.; DADIZADEH, S. A survey on cloud computing. **University of British Columbia Technical Report for CS**, v. 508, p. 55–58, 2009.

GREENBERGER, M. **The Computers of Tomorrow**. 1964. Disponível em: http://www.tnellen.com/cybereng/ebooks/greenbf.htm.

GRIFFIN, R. **Internet Governance**. ETP, 2018. (Internet Governance). ISBN 9781788823548. Disponível em: https://books.google.com.br/books?id=qPw8wAEACAAJ.

GURU99. **Kubernetes Tutorial for Beginners: Basics, Features, Architecture**. 2021. Disponível em: https://www.guru99.com/kubernetes-tutorial.html.

IBM. **IBM System/360 Operating Sytem**. [S.l.], 1968.

Kassem, G. et al. Tcp variants: An overview. In: **2010 Second International Conference on Computational Intelligence, Modelling and Simulation**. [S.l.: s.n.], 2010. p. 536–540.

KEDEM, A. B. M. K. Z.; WYCKOFF, P. Charlotte: Metacomputing on the web. In: CITESEER. **In The 9th ICPDCS International Conference on Parallel and Distributed Computing and Systems**. [S.l.], 1996.

KLEINROCK, L. Information flow in large communication nets. **RLE Quarterly Progress Report**, v. 1, 1961.

KLEINROCK, L. **Communication nets: Stochastic message flow and delay**. [S.l.]: Mcgraw-Hill, 1964.

KUBERNETES. **First commit**. 2014. Disponível em: https://github.com/kubernetes/kubernetes/commit/2c4b3a562ce34cddc3f8218a2c4d11c7310e6d56.

KUROSE, J. F. **Computer networking: A top-down approach featuring the internet, 3/E**. [S.l.]: Pearson Education India, 2005.

LEINER, B. M. et al. A brief history of the internet. **ACM SIGCOMM Computer Communication Review**, ACM New York, NY, USA, v. 39, n. 5, p. 22–31, 2009.

LIU, F. et al. Nist cloud computing reference architecture. **NIST special publication**, v. 500, n. 2011, p. 1–28, 2011.

MARILL, T.; ROBERTS, L. G. Toward a cooperative network of time-shared computers. In: **Proceedings of the November 7-10, 1966, fall joint computer conference**. [S.l.: s.n.], 1966. p. 425–431.

MARSHALL, M. **Google acquires online word processor, Writely**. 2006. Disponível em: https://venturebeat.com/2006/03/09/google-acquires-online-word-processor-writely/.

MAYFIELD, D. 2016. Disponível em: https://www.gosquared.com/blog/saas-welcome-email-examples.

MCDONALD, P. **Introducing Google App Engine + our new blog**. 2008. Disponível em: http://googleappengine.blogspot.com/2008/04/introducing-google-app-engine-our-new.html.

MCLELLAN, C. **Cloud computing in the real world: The challenges and opportunities of multicloud**. 2021. Disponível em: https://www.zdnet.com/article/cloud-computing-in-the-real-world-the-challenges-and-opportunities-of-multicloud/.

MELL, P.; GRANCE, T. et al. The nist definition of cloud computing. Computer Security Division, Information Technology Laboratory, National ..., 2011.

MICROSOFT. **Hyper-V RTM announcement. Available today from the Microsoft Download Centre**. 2008. Disponível em: https://web.archive.org/web/20080629055418/http://blogs.technet.com/jhoward/archive/2008/06/26/hyper-v-rtm-announcement-available-today-from-the-microsoft-download-centre.aspx.

MICROSOFT. **Microsoft 365**. 2021. Disponível em: https://www.microsoft.com/en-us/microsoft-365.

MURRAY-WEST, R. **From 1876 to today: how the UK got connected**. 2016. Disponível em: https://www.telegraph.co.uk/technology/connecting-britain/timeline-how-uk-got-connected/.

NETFLIX. **About Netflix**. 2021. Disponível em: https://about.netflix.com/en.

NYANSA. 2016. Disponível em: https://www.nyansa.com/news/press-releases/youtube-netflix-microsoft-office-365-top-list-best-performing-saas-applications-used-corporate-network

PENG, J. et al. Comparison of several cloud computing platforms. In: IEEE. **2009 Second international symposium on information science and engineering**. [S.l.], 2009. p. 23–27.

PRAJAPATI, A. G.; SHARMA, S. J.; BADGUJAR, V. S. All about cloud: a systematic survey. In: IEEE. **2018 International Conference on Smart City and Emerging Technology (ICSCET)**. [S.l.], 2018. p. 1–6.

QTS. 2016. Disponível em: https://www.qtsdatacenters.com/resources/articles/introducing-healthcare-community-cloud.

REDHAT. **Types of cloud computing**. 2021. Disponível em: https://www.redhat.com/en/topics/cloud-computing/public-cloud-vs-private-cloud-and-hybrid-cloud.

REDHAT. **Understanding OpenStack**. 2021. Disponível em: https://www.redhat.com/en/topics/openstack.

REDHAT. **What is hybrid cloud?** 2021. Disponível em: https://www.redhat.com/en/topics/cloud-computing/what-is-hybrid-cloud.

REDHAT. **What is virtualization?** 2021. Disponível em: https://www.redhat.com/en/topics/virtualization/what-is-virtualization.

REDHAT. **What's the difference between cloud and virtualization?** 2021. Disponível em: https://www.redhat.com/en/topics/cloud-computing/cloud-vs-virtualization.

RENO, N. **Cloud Market Ends 2020 on a High while Microsoft Continues to Gain Ground on Amazon**. 2021. Disponível em: https://www.srgresearch.com/articles/cloud-market-ends-2020-high-while-microsoft-continues-gain-ground-amazon.

RIMAL, B. P.; CHOI, E.; LUMB, I. A taxonomy and survey of cloud computing systems. In: IEEE. **2009 Fifth International Joint Conference on INC, IMS and IDC**. [S.l.], 2009. p. 44–51.

ROBERTS, L. G. Multiple computer networks and intercomputer communication. In: **Proceedings of the first ACM symposium on Operating System Principles**. [S.l.: s.n.], 1967. p. 3–1.

RUPARELIA, N. B. **Cloud computing**. [S.l.]: Mit Press, 2016.

SALESFORCE. 2021. Disponível em: https://www.salesforce.com/company/our-story/.

SALESFORCE. **What is Salesforce?** 2021. Disponível em: https://www.salesforce.com/products/what-is-salesforce/.

SAREEN, P. Cloud computing: types, architecture, applications, concerns, virtualization and role of it governance in cloud. **International Journal of Advanced Research in Computer Science and Software Engineering**, v. 3, n. 3, 2013.

SHOPIFY. 2021. Disponível em: https://www.shopify.com/.

SRIVASTAVA, A. **Introducing Windows Azure**. 2010. Disponível em: https://web.archive.org/web/20100514093158/http://blogs.msdn.com/windowsazure/archive/2008/10/27/introducing-windows-azure.aspx.

STATISTA. **Global digital population as of January 2021**. 2021. Disponível em: https://www.statista.com/statistics/617136/digital-population-worldwide/.

STATISTA. **Public cloud services end-user spending worldwide from 2017 to 2022 (in billion U.S. dollars)**. 2021. Disponível em: https://www.statista.com/statistics/273818/global-revenue-generated-with-cloud-computing-since-2009/.

STEEN, M. v. **Possible mistake at Distributed Systems 3rd edition**. 2021. Não publicado. Conversação em e-mail.

SULLIVAN, D. **What Legal Recourse Against Cloud Services Do You Have?** 2017. Disponível em: https://www.cloudwards.net/what-legal-recourse-against-cloud-services-do-you-have/.

SUSANTA, N.; TZI-CKER, C. A survey on virtualization technologies. **Experimental Computer Systems Lab**, 2005.

TANENBAUM, A. S.; STEEN, M. V. **Distributed systems: principles and paradigms**. [S.l.]: Prentice-hall, 2007.

THOUSANDEYES. **Cloud Performance Benchmark**. [S.l.], 2019–2020.

UCL. **UCL engineers set new world record internet speed**. 2020. Disponível em: https://www.ucl.ac.uk/news/2020/aug/ucl-engineers-set-new-world-record-internet-speed.

VLADIMIRSKIY, V. 2016. Disponível em: https://getnerdio.com/academy/10-popular-software-service-examples/.

VMWARE. **VMware Timeline**. 2021. Disponível em: https://www.vmware.com/timeline.html.

WCG. 2021. Disponível em: https://www.worldcommunitygrid.org/.

WIKIPEDIA. 2021. Disponível em: https://en.wikipedia.org/wiki/Multitenancy.

WIKIPEDIA. 2021. Disponível em: https://en.wikipedia.org/wiki/Mobile_backend_as_a_service.

WIKIPEDIA. **Docker (Software)**. 2021. Disponível em: https://en.wikipedia.org/wiki/Docker_(software).

WIKIPEDIA. **Edge computing**. 2021. Disponível em: https://en.wikipedia.org/wiki/Edge_computing.

WIKIPEDIA. **Infrastructure as Code**. 2021. Disponível em: https://en.wikipedia.org/wiki/Infrastructure_as_code.

WIKIPEDIA. **LXC**. 2021. Disponível em: https://en.wikipedia.org/wiki/LXC.

WIKIPEDIA. **Pool (Computer Science)**. 2021. Disponível em: https://en.wikipedia.org/wiki/Pool_(computer_science).

WIKIPEDIA. **VirtualBox**. 2021. Disponível em: https://en.wikipedia.org/wiki/VirtualBox#History.

WORLDCOMMUNITYGRID. **OpenPandemics - COVID-19**. 2021. Disponível em: https://www.worldcommunitygrid.org/research/opn1/overview.do.

YI, S. **SQL Azure - The Year in Review**. 2011. Disponível em: https://azure.microsoft.com/en-us/blog/sql-azure-the-year-in-review/.

YOUTUBE. **Sobre o YouTube**. 2021. Disponível em: https://www.youtube.com/about/.

ZHOU, S. et al. Utopia: a load sharing facility for large, heterogeneous distributed computer systems. **Software: practice and Experience**, Wiley Online Library, v. 23, n. 12, p. 1305–1336, 1993.